This is an accepted author manuscript of an article published by Elsevier in
Games and Economic Behavior on 13 March, 2013, available online
http://dx.doi.org/10.1016/j.geb.2013.03.002

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

Topi Miettinen†

February 8, 2013

Abstract

I analyze how informal agreements can be sustained by moral emotions with regard to a large class of two-player games. Specifically, I assume that people feel guilty if they breach an agreement and that the guilt increases according to the degree of the harm inflicted on the other. A central insight is that it is easier to sustain efficient informal agreements if actions are strategic complements than if they are strategic substitutes. I complement this general insight by studying two specific cases where negotiators face uncertainty about the breach of the agreement. I show that while the optimal agreement in a game with strategic substitutes must compromise on surplus-maximization and efficiency, the optimal agreement in a game with sufficiently strong strategic complements tends to maximize both the surplus and the probability of compliance especially if the game is symmetric.

JEL Classification: C72, C78, D03

Keywords: partnerships; contracts; pre-play communication; social norms; guilt

*Previous versions circulated with titles "Promises and Conventions - A Theory of Pre-play Agreements" and "Contracts and Promises - An Approach to Pre-play Agreements".

†SITE, Stockholm School of Economics & Hanken School of Economics at HECER, P.O.Box 479, Fi-00101 Helsinki; topi.miettinen@hanken.fi; +358403521406

1 Introduction

In this paper I consider a fairly general two-player model which studies the effectiveness of informal agreements in promoting efficiency. In the model the parties to an agreement feel guilty when breaching their promise. I assume that the guilt cost increases with the harm the breach inflicts on the other.[1] Without an agreement, however, the players neglect the externality on their partner. An agreement may thus promote the alignment of interests between the two parties.

I show that whether an informal Pareto-efficient agreement can be struck, subtly hinges upon the degree of strategic complementarity in the underlying interaction.[2] There is a conflict between the Pareto-efficiency of the agreement and the incentive to abide by the agreement when actions are *strategic substitutes*: making the terms of the agreement more efficient necessarily weakens the incentives to abide by the deal. With sufficiently strong strategic complements such a conflict is circumvented: if an agreement requires players to play a strategy profile, where actions are symmetrically increased above the underlying game equilibrium level, the agreement will not be violated. There also exists a Pareto-efficient agreement that will not be breached. This holds even when strategic complementarity is so weak that no efficient underlying game equilibrium exists. I also study two specific cases where negotiators face uncertainty about the breach of the agreement. First, I consider negotiators wishing minimize the probability of breach when striking an agreement. I show that when negotiators have logit-quantal-response trembling hands and probability of trembling approaches zero, the optimal agreement approaches an underlying game equilibrium when actions are strategic substitutes. Thus preplay-agreements deliver no efficiency gains to guilt-averse players. To the contrary when actions are sufficiently strong strategic complements, the agreement approaches the upper boundary of the strategy space which also constitutes the set of Pareto-efficient profiles. Second, if the proneness to guilt types are unknown when the agreement is negotiated and the uncertainty is lifted before the actions are chosen, the agreement which maximizes the expected surplus will compromise on ex-post efficiency when actions are strategic substitutes. Yet with sufficiently strong strategic complements, the optimal agreement tends to maximize the ex-post surplus and minimize the risk of breach especially when the game is symmetric.

Key findings in experimental economics indicate that a dislike of breaching informal agreements warrants careful theoretical study: First, allowing subjects to communicate without giving them opportunity to write enforceable contracts increases contributions in public good games (Ledyard, 1995; Sally, 1995). Second, people prefer not to lie or let down others (Gneezy, 2005;

[1] An assumption in line with the empirical finding of Gneezy (2005) that willingness to lie is inversely related to the cost that the lie inflicts on the other.

[2] Actions are strategic complements if the incentive to increase one's action increases in the action of the other; the opposite holds with strategic substitutes. See Vives (2005) for an introduction to games with strategic complementarities.

Charness and Dufwenberg, 2006, 2011; Vanberg, 2008; Sutter, 2009; Erat and Gneezy, 2011).[3] In public good games, agreeing to contribute more than one actually intends to contribute is a lie which harms others. Thus, the dislike of lying if it harms others might help to better understand the increased contributions in public good games. This paper contributes to this goal by proposing a formal framework and identifying strategic complementarity as a key factor for the success of informal agreements to promote Pareto-efficiency.

Results from communication experiments lend some support to the theoretical findings. Suetens (2005) found that allowing for communication induced higher rates of cooperation in an R&D environment which is characterized by strategic complements but that there is no cooperation in an environment with strategic substitutes. Isaac and Walker (1988) looked at a voluntary provision public good game with a constant-returns-to-scale technology that implied a setting with weak strategic complements. They found that communication had a strong positive effect on efficiency. The experimental setup induced practically first-best efficient average contribution levels. In a design with decreasing returns to scale, implying strict strategic substitutes, they found that communication increased contributions much less. Isaac, McCue, and Plott's (1985) similar finding in a design with decreasing returns to scale further supports this. In this paper I contribute with an explanation for why efficient agreements tend not to be abided by when actions are strategic substitutes and why they tend to be complied with when actions are strategic complements. I also provide an explanation why the efficient agreements might be agreed upon in game where actions are strategic complements and not in games where they are strategic substitutes: Pareto-efficiency is in line with the probability that neither breaches the agreement in the former but conflicting in the latter type of games.

The closest study to the present one is the independent, mainly experimental contribution by Kessler and Leider (2012) which examines informal agreements with four alternative underlying games. They find that in games with strategic complements informal agreements are more successful in promoting higher surplus than in games without strategic complements. They also show that this surplus gap widens with (random strangers) repetition. The explanation they suggest hinges on assumptions of incomplete information and bounded rationality and their line of argument is in parellel with the equilibrium and bounded rationality comparative statics analyses of Athey (2001), Vives and van Zandt (2007), and Schmutzler (2011): when actions are strategic complements and the opponent's action is raised for intrinsic reasons (for instance, guilt aversion to breach an agreement with maximal actions) also the parties, who have little intrinsic motivation to abide by the agreement, have an incentive to shift their action upwards. My line of argument is different and shows among other things that the intrinsic motivation to abide by the agreement itself is stronger for a Pareto-efficient agreement than for a Pareto-inefficient underlying game equilibrium agreement.

[3]This view is also supported by Hoffman (1982), for instance, who suggests that guilt has its roots in a distress response to the suffering of others which reflects internalized social norms.

Strategic complementarity and implicit contracts play a central role in industrial organization, labor economics, and organizational economics, for instance. Small businesses, such as retail shops, provide a good example. They are often located close to each other allowing for frequent informal relationships and communication. Yet, explicit contracting or collusion between competing businesses may be prohibited by law. Informal agreements may be the only way to sustain price cartels. To understand whether and when informal agreements can help to sustain collusion, one needs to understand how social and psychological principles factor into and interact with the pecuniary incentives of the underlying interaction. Another core application are organizations where complementarities play a central role and explicit informal agreements or implicit ones coded in the corporate culture are central for successful cooperation in the workplace. The present model contributes to understanding why collusion or cooperation may be more prevalent in contexts with strategic complements than with substitutes.

My model contributes to the large and growing theoretical literature analyzing the effects of agreements and communication on interaction in games. Building upon Farrell (1987, 1988), there is a literature on pre-play communication of intentions.[4] I depart from this literature by allowing deviations from pre-play messages or agreements to be costly.[5] I assume these costs to satisfy the principle of more guilt for greater harm.[6] Guilt has been discussed in several papers since Frank (1988) argued that it may well be materially profitable for an agent to have a conscience - a dislike for disobeying social norms. Kandel and Lazear (1992) study a model of guilt and shame in partnership situations but they do not address strategic complementarity. More specific and experimentally motivated models are proposed by Ellingsen and Johannesson (2004), Bicchieri (2006), Lopez-Perez (2008, 2012), Charness and Dufwenberg (2006), and Battigalli and Dufwenberg (2007).[7] The first four papers lay out more traditional outcome-based models where the agreement is considered an argument in the utility function and thus part of the outcome of the game. The agreement is interpreted explicitly as an informal agreement or tacitly as a social norm, and guilt is felt about transgressing the agreement or the norm. The latter two papers contrive models of aversion for letting down the expectations of the other. The role of the agreement is to shift the expectations in the direction of the agreement implying higher costs of choosing an action further away from the agreed action. Since the expectations are arguments in the utility function, these papers by definition fall into the category of psychological games (Geanokoplos et al., 1989; Battigalli and Dufwenberg, 2009). I extend the existing

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

[5]See Demichelis and Weibull (2008) mentioned above or Ellingsen and Östling (2010) and Crawford (2003) where players are boundedly rational and some players always prefer sticking to their pre-play promises.

[6]In sociology and social psychology this principle is called the just deserts principle (Hamilton and Rytina, 1980; Darley et al., 2000; Carlsmith et al., 2002).

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

outcome-based models of guilt by essentially allowing guilt to be increasing in the inflicted harm.[8] This feature is crucial to the results. My model also has a straightforward psychological game interpretation. In that interpretation the focus is on whether any player has an incentive to deviate from an agreement if she believes that others will abide by the agreement. If not, abiding by the agreement is a psychological Nash equilibrium.

The paper is organized as follows. Section 2 presents the model and Section 3 illustrates the implications for voluntary public good provision. In Section 4, the main results are presented and in Section 5 the model is applied to a Cournot-duopoly with product differentiation. Section 6 concludes.

2 The model

2.1 The underlying game

For simplicity, I limit the analysis to two-player games. The point of departure for the analysis is the basic underlying game Γ prior to which the players may strike an agreement m on how to play. Γ is defined by the set of players $i = 1, 2$, the finite action sets of the players S_i, and the underlying game payoff functions $\pi_i : S \to R$ which can be thought of mapping the outcomes, $s = (s_1, s_2) \in S = S_1 \times S_2$, to monetary consequences. I mainly focus on finite games with ordered strategies and, without having to narrow focus to any subclass of such games, I label the actions from 0 to n_i, $S_i = \{0, ..., n_i\}$ where I call n_i the maximal action of player i.

For any given action of player i, her payoff is increasing in the action of the other player, j, i.e. the marginal external return is always positive. It is thus natural to call the actions contributions, inputs, or efforts. Notice that this assumption is not as restrictive as it first seems: Once it is assumed that payoffs are either both strictly increasing or both strictly decreasing in the other player's action, assuming the former is without a loss of generality. If both payoffs are decreasing in the opponent's action, we can make the payoffs increasing again by reversing the ordering of each strategy set. This has no effect on the second differences. Thus, games with decreasing payoffs in the other player's action can be analyzed using the same artillery.[9]

I also assume that there are decreasing marginal private and external returns to increasing one's action in the sense that payoff is concave (weakly) in own action and in that of the other player. By parameters δ_i and σ_j , I denote the rate of decay of the marginal private and external return, respectively. By concavity, for all s,

$$\delta_i = \pi_i(s_i + 1, s_j) - \pi_i(s_i, s_j) - [\pi_i(s_i, s_j) - \pi_i(s_i - 1, s_j)] \leq 0, \tag{1}$$

[8] In Bicchieri's model guilt depends linearly on the inflicted harm while the other two models do not let guilt to depend on the inflicted harm.

[9] I will illustrate this in Section 5.

and

$$\sigma_i = \pi_i(s_i, s_j + 1) - \pi_i(s_i, s_j) - [\pi_i(s_i, s_j) - \pi_i(s_i, s_j - 1)] \leq 0. \quad (2)$$

Ultimately with sufficiently high action, the marginal private return may turn negative.

I further assume that actions are either strategic substitutes or complements. Actions are strategic substitutes (complements) if i's incentive to increase her action declines (rises) in the opponent's action. Formally, for all $s \in S_1 \setminus \{0\} \times S_2 \setminus \{0\}$,

$$\pi_i(s_i, s_j) - \pi_i(s_i - 1, s_j) < \pi_i(s_i, s_j - 1) - \pi_i(s_i - 1, s_j - 1) \quad (3)$$

$(\pi_i(s_i, s_j) - \pi_i(s_i - 1, s_j) \geq \pi_i(s_i, s_j - 1) - \pi_i(s_i - 1, s_j - 1))$. Let us denote this effect of j's action on the marginal payoff of i by ϕ_i which is assumed constant in the entire set of outcomes.[10] These assumptions ensure that a Nash equilibrium exists in the underlying game.[11] I denote the underlying game best-reply correspondence of player i by $BR_i : S_j \longrightarrow 2^{S_i}$, and the set of i's best-replies to s_j by $BR_i(s_j)$. I restrict attention to games with pure strategy Nash equilibria.[12]

2.2 Agreement

Before the game is played the players can enter into an agreement $m = (m_1, m_2) \in S$. If only player i deviates from the agreed action m_i and plays a feasible action $s_i \neq m_i$ then

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

Notice that the primary interest is on downward deviations since upward deviations inflict no harm given that payoffs are strictly increasing in the opponent's action.

2.3 Guilt

Players are prone to guilt. If there is an agreement in place, each (generically) feels bad about not fulfilling her part of the deal. Player i's guilt cost, $\theta_i max\{h_j(m, s_i), 0\}^{\varphi}$, is continuous,

[10] ϕ_i is defined as the difference between the left and the right hand side of inequality (3).

[11] One can apply Topkis (1979) which exploits Tarski's fixed point theorem.

[12] Formally, an extension to mixed strategies could easily be done. Yet, the extension involves challenging modelling choices which are peripherical for the main ideas in this paper: if the supports of a player's deviant mixed strategy and her agreed mixed strategy overlap but do not coincide, the outcome may or may not allow the opponent to infer with certainty whether the player has deviated. Whether guilt is simply a function of the expected harm inflicted or whether it depends on the opportunities for public blame mediated by the strength of public signals for deviations, is an interesting question with possibly important implications but less so for the focus in this paper. Battigalli and Dufwenberg (2007) and Sebald (2010) provide insightful analyses of highly related issues.

non-negative and increasing in the inflicted harm. Moreover, I assume that $\varphi \geq 1$ so that guilt is convex in the inflicted harm and that the utility function in the entire game is additively separable in guilt and the underlying game payoff.

$$U_i(m,s) = \begin{cases} \pi_i(s) - \theta_i max\{h_j(m,s_i),0\}^{\varphi} & if\ s_j = m_j \\ \pi_i(s) & otherwise. \end{cases} \quad (4)$$

The entire game payoff depends on m and, due to guilt, it is costly to deviate from the agreement.[13] The parameters $\theta = (\theta_1, \theta_2)$ capture the players' proneness to guilt. For a given inflicted harm, a player with higher proneness to guilt suffers a higher cost. I only allow for non-negative proneness to guilt, $\theta_i \in [0, \infty)$. Non-negative guilt cost and non-negative θ rule out revengeful feelings or spite, on the one hand, and positive emotions related to respecting agreements, on the other hand. I suppose that the proneness to guilt types are common knowledge. This is a restrictive but simplifying assumption. I discuss the implications of relaxing this assumption in the concluding section.

The assumptions imply that if an agent inflicts strictly positive harm on a complying partner, then the guilt cost is strictly positive. Otherwise it is zero. Very small deviations only induce very little guilt while larger deviations matter more, a natural feature not incorporated in constant guilt cost models (Ellingsen and Johannesson, 2004; Lopez-Perez, 2008, 2012). Convexity of guilt in the inflicted harm is harder to justify than continuity and calls for further empirical scrutiny, but I will adopt it since it facilitates analysis.

Remark also, that i's guilt cost depends on the agreed action of i, m_i, only indirectly through the harm. In addition, choosing the agreed action minimizes the guilt cost. There is no guilt if the opponent deviates from the agreement.[14] If this assumption did not hold, one player could marginally deviate from the agreement while the other would comply and both parties might still prefer making an agreement to not making one. While this plausible assumption is not essential for the results de facto, Lopez-Perez (2008, 2012) and Bicchieri (2006, ch. 1) propose game-theoretic models of social norms which make similar assumptions. Bicchieri discusses in length why conditional conformism captured by this assumption is necessary in any reasonable account of social norms:

'Imagine a situation in which someone expects others not to conform to a norm of truth telling. He has observed people openly lying and has been lied to often enough to expect further dishonesty. Yet, he is made to believe that he is expected to conform to a norm of truth-telling. It is likely that this individual would consider the expectation illegitimate and feel no guilt at violating it' (Biccieri, 2006, pp. 24)

In the main text I shall further constrain interest in guilt cost which is linear in the inflicted

[13] The results of the paper would hold with infinite action sets if payoffs in Γ are twice continuously differentiable and the right hand derivative of the guilt cost w.r.t. harm is strictly positive at zero (as in the case with $\varphi = 1$). See Section 5.

[14] Miettinen and Suetens (2008) provide some further empirical support for this assumption.

harm whenever harm is positive and proceed by imposing $\varphi = 1$. The proofs in the appendix are carried out with the more general formulation.

The utility model put forward in this paper is most closely related to the model of promises by Ellingsen and Johannesson's (2004), and Bicchieri's (2006), and Lopez-Perez's (2008) models of social norms. Bicchieri and Lopez-Perez allow for an arbitrary number of players and in that sense their models are more general than the present one. Social norms that they study are closely related to pre-play agreements. Social norms guide agreements; sometimes an agreement merely explicitly restates and confirms the code of conduct already implicitly agreed upon in a social norm. The main difference between Bicchieri's and Lopez-Perez's models is that in the latter the harm inflicted on others bears no influence on the extent of guilt while Bicchieri assumes a linear association between guilt and harm. The model in the main text of the present paper is thus essentially a two-player version of Bicchieri's model applied to the pre-play agreement framework. In the appendix I generalize the model allowing for a strictly convex guilt cost in the inflicted harm.

2.4 Incentive compatible agreements

Suppose the players have agreed to play m. Let $\Gamma(m;\theta)$ denote the ensuing non-cooperative game. The net payoff of player i who has agreed on m can be written as $U_i : S \times S \longrightarrow \mathbb{R}$, where

$$U_i(m,s) = \begin{cases} \pi_i(m) + g_i(m,s_i) - \theta_i max\{h_j(m,s_i),0\} & if\ \ s_j = m_j, \\ \pi_i(s) & otherwise. \end{cases} \tag{5}$$

The payoff to player j who complies with the agreement can be written as $U_j(m_j,m_i,m_j,s_i) = \pi_j(m) - h_j(m,s_i)$. Once having entered into an agreement, each player trades off the gains and the guilt cost from breaching; this is assuming that the other player does not deviate. An agreement is *incentive compatible for* i if for all $s_i \in S_i$

$$g_i(m,s_i) \leq \theta_i max\{h_j(m,s_i),0\}. \tag{6}$$

I assume that a player, who is indifferent to breaking or keeping her promise, abides by the agreement.[15] When the incentive compatibility condition holds for both players, the agreement m is a Nash equilibrium of the game, $\Gamma(m;\theta)$. Guilt cost can only strengthen the incentives to play a given profile of actions. Hence, the following holds.

Remark 1 *If the agreement m is a Nash equilibrium of the underlying game, then it is incentive compatible for i for any proneness to guilt type $\theta_i \in [0,\infty)$.*

On the other hand, when neither player is intrinsically motivated, the underlying game payoffs characterize the incentives and only Nash equilibria can be agreed upon.

[15] Demichelis and Weibull (2008) study pre-play communication under such an assumption and show that Nash equilibria are evolutionarily stable if and only if Pareto-efficient.

Remark 2 *If $\theta_i = 0$, then the agreement is incentive compatible for i if and only if $m_i \in BR_i(m_j)$.*

Finally, since intrinsic incentives evaporate where the opponent deviates, the following holds.

Remark 3 *Let s^* be an underlying game Nash equilibrium. If $s_i^* \neq m_i$ for $i = 1, 2$, then s^* is also a Nash equilibrium of $\Gamma(m; \theta)$ for all θ.*

Remark 3 clarifies that the model generically has multiple equilibria and the studies helping us to understand the coordinating role of pre-play agreements when multiple equilibria exist (Farrell and Rabin, 1996) are helpful even when one wants to thoroughly understand the commitment role of pre-play agreements. I am interested in the opportunities that promise-making opens for efficient outcomes and thus I focus on the Pareto-efficient equilibria.

Since the opponent's payoff is increasing in one's action, a withdrawal of one promised unit of action inflicts a positive marginal harm (deviation downwards), $\eta_j(m)$. Marginal gain from breaching, $\gamma_i(m)$, can be positive or negative since I do not assume monotonicity in own action.[16] Notice yet that player i will not suffer from guilt if she makes both players better off by deviating. Consequently, i's payoff must be non-increasing in i's action for player i to keep her promise.[17] I denote this feasible set of agreements by M^F. Formally, the feasible set is characterized by

$$M^F = \cap_{i=1,2} M_i^F, \tag{7}$$

where $M_i^F = \{s | \gamma_i(s_i + 1, s_j) \geq 0\}$. Let us first show that, within this set, non-positive marginal incentive to breach, $\gamma_i(m) \leq \theta_i max\{\eta_j(m), 0\}$, is necessary and sufficient for an agreement m to be incentive compatible for i.

Proposition 1 *Let $\sigma_j, \delta_i \leq 0$. Let $m_i \in M_i^F$ and let m_i neither be the maximal nor the minimal action, i.e. $m_i \notin \{0, n_i\}$. Then an agreement m is incentive compatible for i if and only if i's marginal incentive to breach is non-positive,*

$$\gamma_i(m) \leq \theta_i max\{\eta_j(m), 0\}.$$

The simple idea behind this result is as follows. Since harm is the difference between the agreed payoff and the actual payoff, $h_j(m, s_i) = \pi_j(m) - \pi_j(m_j, s_i)$, and payoff is concave in each action, the harm h_j is a convex function of s_i. On the other hand, the underlying game payoff π_i is concave in s_i and, therefore, also the gain from breaching is concave. According to a standard convexity argument, checking that i does not prefer breaching the agreement marginally is necessary and sufficient for an agreement to be incentive compatible for i.[18]

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

[17] Except for the largest action $s_i = n_i$ of course.

[18] In the general case it is crucial that we assume that $\varphi \geq 1$ so that guilt is a weakly convex function in the inflicted harm. Otherwise non-marginal deviations might pay off although a marginal deviation does not.

3 Simple partnership game

Before deriving and discussing the main results in full generality, let us illustrate the implications of the model by means of a public good provision game. 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 $\pi_i(s) = G(\sum_{k=1,2} s_k) + 10 - s_i$ where the production technology $G(.)$ maps the sum of contributions into the produced amount of the public good. I suppose that for all strategy profiles (s_1, s_2), $G'(\sum s) < 1$. Hence, it is a strictly dominant strategy and a Nash equilibrium strategy to contribute nothing. G' is the marginal per capita return (MPCR). Whenever $2G' > 1$, the generated social surplus increases in each contribution and increasing both contributions by the same amount constitutes a Pareto-improvement.

Let the production technology have either constant or decreasing returns to scale, $G'' \leq 0$. If the returns to scale are strictly decreasing, actions are strategic substitutes: the more is contributed by the other, the lower the marginal benefit of a further contribution. If there are constant returns to scale, the game has weak strategic complements: if the opponent contributes more, the marginal benefit of the player's further contribution is unaffected.

Players can strike any agreement where guilt is sufficient to prevent either from breaching. Since the production technology has non-increasing returns to scale, the private and external returns of each contribution are weakly concave. According to Proposition 1, the incentive compatibility conditions are characterized by the simpler marginal incentive conditions. It is thus necessary and sufficient to check for the one dollar underprovision only. Given that j abides by the agreement, player i breaches her promise if and only if

$$\begin{aligned} 1 - G(\textstyle\sum_{k=1,2} m_k) + G(\sum_{k=1,2} m_k - 1) \\ > \theta_i[G(\textstyle\sum m_k) - G(\sum m_k - 1)]. \end{aligned} \tag{8}$$

The term on the upper line is the marginal gain from breaching: one gains back the dollar one decides to withdraw but providing one unit less than promised also leaves some of the public good unproduced. The amount of public good lost by each of the two partners equals $G(\sum_{k=1,2} m_k) - G(\sum_{k=1,2} m_k - 1)$ and it is therefore also the harm one inflicts on one's partner.

A property explicit in (8) is worth emphasizing: if there are decreasing returns to scale, $G'' < 0$, the incentives to abide by the agreement are weaker if one agrees to contribute to a profile which Pareto-improves the underlying game equilibrium than if one agrees on the equilibrium profile itself. The more is produced, the lower the marginal product of the marginal contribution and thus the harm on the other is also lower if the marginal contribution is withdrawn. On the other hand, the deviator gains back her dollar and loses the marginal product: the marginal gain from a unit underprovision vis-à-vis the agreement is $1 - h_j(m, m_i - 1)$ which is increasing in the sum of contributions. Therefore the conflict between efficiency and incentives to abide by the agreement is evident:

Remark 4 *Let $G'' < 0$. If both agreed contributions are increased, then the marginal gain from breaching increases and the marginal harm on the opponent decreases.*

Yet, if there are constant returns to scale, $G' = \alpha$, the marginal payoffs are constant and the marginal breach incentive condition reduces to

$$(1 - \alpha) \geq \theta_i \alpha.$$

Any changes in the agreed actions do not affect the breaching incentives insofar as the underlying game dominant strategy is not agreed upon. Hence, with constant returns to scale, the incentives and efficiency are not in conflict: if an agreement to play some underlying game disequilibrium strategy profile is kept, then an agreement to play an efficient profile will be kept.

Remark 5 *Let G' be constant. If $m \in S$ is such that for $i = 1, 2$ the agreement on m is incentive compatible and each agreed contribution is positive, $m_i > 0$, then there is an efficient agreement which is incentive compatible for $i = 1, 2$.*

Proposition 2 summarizes the findings of this section.

Proposition 2 *In the public good game,*

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

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

(c) *if $G' = \alpha$, player i's marginal incentive to breach is decreasing in α and constant in m_j and m_i.*

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

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 above a unique underlying game equilibrium 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).

Furthermore in regards to the experiments by Isaac, McCue and Plott (1985) and Isaac and Walker (1988), the results suggest that the reason for lower cooperation rates in interior group optimum designs may not be the difficulty of identifying and agreeing on an interior group optimum, as conjectured by Isaac and Walker. Rather the incentives to respect agreements when sufficiently close to an interior group optimum are very weak in their experiment, whereas the incentives are the strongest when close to a boundary group optimum. In order to account for this difference, it is crucial that the guilt cost is linear (or weakly convex) in the inflicted harm.

If the guilt cost is constant conditional on inflicting harm[19], then guilt is concave in the harm on the other unconditionally. A constant guilt cost induces a payoff-discontinuity that provides a strong incentive for not deviating marginally from the agreement but provides no incentive for preventing further marginal deviations.[20] Therefore the constant-guilt-cost approach cannot account for the differences in the interior and boundary group optimum experiments in a satisfactory manner.

In Section 4, I generalize the sharp contrast between the constant returns to scale technology and the decreasing returns to scale technology in public good production: I shall show that there is a conflict between incentives and Pareto-efficiency when actions are strategic substitutes and the payoffs are same-sign-monotone in the opponent's action. A sufficiently strong strategic complementarity resolves any such conflict and any Pareto-improving profile is incentive compatible when players have a positive proneness to guilt in the first place. I shall also consider negotiators attempting to choose an optimal agreement when there is uncertainty about breach. I will show that such negotiators tend to choose an ex-post Pareto-efficient agreement when actions are strategic complements but must compromise on efficiency to reduce the likelyhood of breach when actions are strategic substitutes.

4 Analysis

4.1 Incentive compatible agreements and Pareto-efficiency

Increasing one's action results in a private return to oneself and an external return to one's partner. The marginal private and external return are decreasing at a rate captured by δ_i and σ_j, respectively. While the marginal external return is globally positive – for all s, $\pi_j(s_j, s_i + 1) - \pi_j(s_j, s_i) > 0$ – the private return is positive only up to the best-reply contribution and negative thereafter. The best-reply correspondences are either weakly increasing (strategic complements) or decreasing (strategic substitutes) and, given the constant second differences, the approximate slope of the correspondences is constant $-\phi_i/\delta_i$. In typical cases (when the complementarity is not very strong), there exists a unique interior equilibrium where these best-reply-correspondences meet.

From a welfare perspective, the best profile that the players can reach non-cooperatively without striking an agreement, is the Pareto-dominating equilibrium when it exists (strategic complements) or one of the many equilibria which cannot be Pareto-ordered (strategic substitutes).[21] When strategic complementarity is not strong or actions are strategic substitutes, this

[19] As is assumed in Ellingsen and Johannesson (2004), for instance.

[20] A discrete function defined on non-negative real numbers, which takes a positive constant value unless the argument equals zero in which case the function takes value zero, is clearly concave.

[21] Milgrom and Roberts (1990) show that equilibria can be Pareto-ranked when actions are strategic complements and payoffs are increasing in the opponent's action.

equilibrium is interior and/or inefficient. Let us denote by s^* the underlying game equilibrium on which the players would coordinate without an agreement.[22]

In striking an informal cooperative agreement, players strive to reach a Pareto-improvement to the underlying game equilibrium s^*. Increasing a player's action marginally when the current agreement is m changes the social surplus by

$$\begin{aligned}\pi_i(m_i+1,m_j)-\pi_i(m_i,m_j)+\pi_j(m_j,m_i+1)-\pi_j(m_j,m_i)\\ =-\gamma_i(s_i^*+1,s_j^*)+\eta_j(s_i^*+1,s_j^*)\\ +(m_i-s_i^*)(\delta_i+\sigma_j)\\ +(m_j-s_j^*)(\phi_i+\phi_j)\end{aligned} \tag{9}$$

where I have used the fundamental theorem of calculus to derive the expression on the right-hand side of the equality. In particular, if the status quo agreement is the underlying game equilibrium s^* then $m_i - s_i^* = 0 = m_j - s_j^*$ and $\gamma_i(s_i^*+1,s_j^*) \approx 0$ while $\eta_j(s_i^*+1,s_j^*) > 0$ for $i,j = 1,2$. Therefore, social welfare dictates that i's action should be increased at a level above the equilibrium. At an interior surplus maximizing action profile the expression in equation (9) is (approximately) zero for both players' actions simultaneously, and by the above argument the surplus maximizing profile lies to the north-east of the interior equilibrium s^*.

Pareto-improvements to an interior equilibrium of Γ may be reached by increasing both actions simultaneously. As one attempts to promote Pareto-efficiency by increasing actions, the declining private and external marginal return pose a challenge. The declining marginal returns tend to magnify the private gain and lessen the harm of deviating marginally from the agreement. These two effects tend to undermine the incentives to abide by the agreement and they are formally stated on the first and the second line of Lemma 1.

Lemma 1 *The marginal gain from breaching is increasing and the marginal harm is decreasing in own agreed action. The marginal gain from breaching is increasing and the marginal harm is decreasing in the agreed action of the opponent if and only if actions are strict strategic substitutes; formally,*

$$\begin{aligned}\gamma_i(m_i+1,m_j)-\gamma_i(m_i,m_j)&=-\delta_i\\ \eta_j(m_i+1,m_j)-\eta_j(m_i,m_j)&=\sigma_j\\ \gamma_i(m_i,m_j+1)-\gamma_i(m_i,m_j)&=-\phi_i\\ \eta_j(m_i,m_j+1)-\eta_j(m_i,m_j)&=\phi_j.\end{aligned} \tag{10}$$

When a Pareto-improvement to an equilibrium of Γ is agreed upon, the other player's promised action is also higher and bears an influence on the player's marginal gain and the harm. The effects of the opponent's agreed action on the marginal private and external effect are formally

[22]Existing literature extensively studies how and whether pre-play communication can achieve coordination when there are multiple equilibria (Farrell and Rabin, 1996). The present paper is concerned on the commitment function of pre-play agreements; we will abstract from the coordination issues.

stated on the third and fourth lines of Lemma 1. These effects are fully determined by strategic complementarity.[23] The intuition for the association between ϕ_i and the private return is standard: when actions are strategic substitutes, a party's marginal private return to higher input is smaller the higher the input of her partner. Thus breaching a promise marginally is all the more tempting as the partner's action is higher. On top of this usual effect, strategic substitutability factors in through an additional channel: there is a negative effect of the partner's higher action on the *external* return of one's own action, ϕ_j, eroding the guilt motivation to keep one's promise. In summary, when actions are strategic substitutes, both the effect of one's own promised action and that of the partner's promised action on the marginal private and external return erode the incentive to abide by the agreement when players strike a Pareto-improving pre-play agreement with respect to the underlying equilibrium profile.

Proposition 3 *Let actions be strategic substitutes. If players agree on playing a strategy profile which Pareto-dominates an interior underlying game equilibrium, s^*, then the marginal gain from breaching is greater and the marginal harm is smaller at the agreed profile than at the underlying game equilibrium.*

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

Let us take a closer look at how the incentive to breach relates to efficiency. Decreasing i's action marginally lowers social surplus if and only if $\gamma_i(m) - \eta_j(m) < 0$. Player i's marginal incentive to breach is negative when $\gamma_i(m) - \theta_i\eta_j(m) < 0$. By imposing $\theta_i = 1$ in the latter condition yields the former condition reflecting the fact that a player with $\theta_i = 1$ fully internalizes the marginal effect on the other and deviates from the agreement if and only if the marginal gain for her is greater than the harm on the opponent. In this case the informal agreement functions much like a Clarke-Groves mechanism guaranteeing that surplus-eroding deviations never take place. If both players' proneness to guilt is above or equal to unity, surplus maximizing agreements can be reached. When θ_i falls short of unity, first-best agreements may be beyond reach.

How much can informal agreements influence welfare when $\theta_i < 1$? To give a precise answer, let us denote by $k_i = m_i - s_i^*$ the number of steps that player i's action at agreement m lies above i's underlying game equilibrium action s_i^*. By using the fundamental theorem of calculus, we can rewrite the marginal harm inflicted on j by i at m as $k_i\sigma_i + k_j\phi_j + \eta_j(s^*)$ and the marginal gain for i as $-k_i\delta_i - k_j\phi_i + \gamma_i(s^*)$. Plugging these expressions into the formal condition in Proposition

[23] One can rearrange inequality (3) so that it captures the effect of higher action of i on the marginal harm inflicted by j: $\phi = \pi_i(s_i, s_j) - \pi_i(s_i, s_j - 1) - [\pi_i(s_i - 1, s_j) - \pi_i(s_i - 1, s_j - 1)] \leq 0$.

1 and solving in terms of θ_i yields

$$\theta_i \geq \frac{\gamma_i(s^*) - k_i\delta_i - k_j\phi_i}{\eta_j(s^*) + k_i\sigma_j + k_j\phi_j} \tag{11}$$

putting a lower bound on θ_i to ensure the incentive compatibility for i at the agreement $m = s^*+k$ (here it is assumed that $\eta_j(s^*) + k_i\sigma_j + k_j\phi_j > 0$; when the opposite holds the agreement is not incentive compatible for any θ_i). Condition (11) clearly expresses the fact that a greater action by the opponent sets a more demanding lower bound for the proneness to guilt of player i when actions are strategic substitutes $\phi_i, \phi_j < 0$.

Alternatively solving for k_i yields

$$k_i \leq \frac{\gamma_i(s^*) - \theta_i\eta_j(s^*) - k_j(\phi_i + \theta_i\phi_j)}{\delta_i + \theta_i\sigma_j} \tag{12}$$

where $\delta_i + \theta_i\sigma_j < 0$ due to decreasing private and external marginal return. Moreover, $\gamma_i(s^*) - \theta\eta_j(s^*)$ can be assumed to be negative given that $\gamma_i(s^*) \approx 0$ and given that we assume $\theta_i > 0$. Condition (12) identifies a boundary for the number of units that i's action can be increased above the equilibrium action so that her marginal incentive to breach is still non-positive. Any agreement $m = s^* + k$ where condition (12) is satisfied for each player $i = 1, 2$ will be kept. The condition indicates that, when actions are strategic substitutes, the higher the action player j promises to carry through, the lower is the greatest feasible action of player i.

Let us analyze a particularly compelling class of agreements: those where incremental actions are equal, $k_i = k_j = k$. Agreements with symmetric increments may be focal or they may appeal to a particular fairness norm. This is the case especially if the underlying game and its equilibrium are symmetric in which case symmetric increments would align with appeals to equate payoffs. When the incremental actions are equal, one can rewrite the incentive condition (12) as follows

$$k \leq \frac{\gamma_i(s^*) - \theta_i\eta_j(s^*)}{\delta_i + \phi_i + \theta_i\sigma_j + \theta_i\phi_j}. \tag{13}$$

Now if actions are strategic substitutes or weak strategic complements, both the denominator and the numerator are negative and there will be an upper limit as to how much promised actions can be increased and still those promises can be kept. Moreover $\theta_i \longrightarrow 0$ implies that the right-hand side approaches zero reflecting the fact that with no proneness to guilt committing to any non-equilibrium profile is prohibitively difficult.

Yet, when actions are strong strategic complements $\phi_i + \theta_i\phi_j > -\delta_i - \theta_i\sigma_j$ then the denominator is positive, the inequality condition (13) is reversed, and any agreement will be kept.[24] This result is of little importance if $\phi_i \geq -\delta_i$ for $i = 1, 2$ in which case there exists a Pareto-efficient

[24]The coefficient of k is positive when solving for k in the marginal incentive compatibility condition. Hence the inequality is reversed in (13) and the right-hand side is negative.

underlying game equilibrium[25] and no intrinsic motivation to abide by an efficient agreement is needed. In contrast when for $i = 1, 2$

$$\phi_i < -\delta_i \tag{14}$$

but

$$\phi_i + \theta_i \phi_j > -\delta_i - \theta_i \sigma_j, \tag{15}$$

there is no Pareto-efficient underlying game equilibrium and informal agreements can result in considerable efficiency gains - even first-best efficient agreements are feasible. A necessary condition for these four conditions to hold is that $\phi_j \geq -\sigma_j$ for $j = 1, 2$. This condition ensures that the opponent's marginally higher promise will raise the marginal harm more (strategic complementarity effect) than the player's own marginally higher promise will lower it (concavity effect): only if the harm on the other rises, one is more inclined to abide by the agreement since one would hurt the other more by breaching one's promise.

Theorem 1 resumes the findings. It also treats asymmetric cases where one player's incentive to abide by the agreement is sustained by underlying game incentives, $m_j \in BR_i(m_j)$, (see Remark 2) while the other's strong proneness to guilt ensures that she will keep her part of the deal, (16). The details are provided in the proof in the appendix which treats a more general case where guilt may be strictly convex in the inflicted harm.

Theorem 1 *Let s^* be an interior Nash equilibrium in the underlying game.*

An agreement $m = s^* + k$ is incentive compatible for i if and only if either $m_i \in BR_i(m_j)$ or

$$k_i \leq \frac{\gamma_i(s^*) - \theta_i \eta_j(s^*) - k_j(\phi_i + \theta_i \phi_j)}{\delta_i + \theta_i \sigma_j}. \tag{16}$$

A Pareto-efficient agreement is incentive compatible for i if either $\phi_i \geq -\delta_i$ or

$$\phi_i + \theta_i \phi_j > -\delta_i - \theta_i \sigma_j.$$

Clearly, there exists a player who will breach an agreement $m + k = (m_1 + k_1, m_2 + k_2)$ if and only if there is $i = 1, 2$ for whom the necessary and sufficient condition stated in the theorem is violated. Likewise, there exists an efficient agreement which neither will breach when the sufficient condition of the theorem is satisfied for $i, j = 1, 2$, $i \neq j$. Notice that the upper bound characterized in (16) is increasing in θ_i when

$$0 < \eta_j(s^*) + k_j \phi_j - \frac{\phi_i}{\delta_i} k_j \sigma_j,$$

which formally states the requirement that player i with zero proneness to guilt will inflict a positive harm on j when best responding to k_j - condition which is known to hold true by the assumption that payoffs are increasing in the action of the other player.[26]

[25]The best-reply correspondences have an approximate slope steeper than one and there will be multiple equilibria, an efficient one at maximal actions. See also Vives (2005).

[26]Notice that I am ignoring the term $\frac{\sigma_j}{\delta_i}\gamma_i(s^*) \approx 0$ due to the fact that $\gamma_i(s^*)$ is known to be approximately zero.

4.2 Choosing the agreement

The above analysis has illustrated that there are potentially many incentive compatible agreements which parties could agree upon. In this subsection I consider two specific set-ups studying which agreement will be chosen by players engaging in pre-play negotiations when there is uncertainty about the probability of breach.

First, I consider negotiators making mistakes in the implementation of their actions and wishing to minimize the probability of breach when striking an agreement. Second, I introduce symmetric incomplete information about proneness to guilt types and assume that the negotiators maximize the expected surplus generated by the agreement (a natural assumption if side-payments are feasible). The ignorance about the proneness to guilt types is lifted once an agreement is in place and before the actions are chosen. In both set-ups I show that the optimal agreement will compromise on ex-post efficiency when actions are strategic substitutes but ex-post Pareto-efficient agreements will be chosen when actions are strategic complements.

The logistic choice model provides a commonly used approach to incorporate mistakes and heterogeneity into decision making. The logit quantal-response equilibrium, for instance, has been fairly successful in grasping many of the empirical patterns.[27] It is based on the idea that actions with higher payoffs are chosen with higher probability and thus the optimal choice should constitute the mode of observed patterns. Let us suppose as in the quantal-response equilibrium that the players make mistakes in the implementation of their strategies and that the mistake probability is inversely related to the opportunity cost of making the mistake. Formally assume that the probability that player i chooses action s_i given that agreement m is in place is given by

$$q_i(s_i|m) = \frac{exp\{\frac{1}{\mu}[\pi_i(m) + g_i(m, s_i) - \theta_i h_j(m, s_i)]\}}{\sum_{s_i' \in S_i}\{\frac{1}{\mu}[exp(\pi_i(m) + g_i(m, s_i') - \theta_i h_j(m, s_i')]\}}.$$

Players knowing that mistakes can take place might prefer agreeing on joint profiles which minimize the risk of breach. To this end, let us consider formally which agreement will be chosen if players wish to choose an agreement which maximizes the probability that the agreement is kept,

$$\hat{m}(\mu) \equiv max_m\{q_1(m_1|m) \times q_2(m_2|m)\}. \tag{17}$$

Alternatively, the players might have an interest in choosing the agreement which minimizes the larger of the breach probabilities of the two players,

$$\tilde{m}(\mu) \equiv max_m min_{1,2}\{q_1(m_1|m), q_2(m_2|m)\}. \tag{18}$$

[27] Anderson et al.(1998), for example, study heterogeneous behavior implied by the logit quantal-response equilibrium in a public good provision setting allowing for altruism. I am simplifying by assuming that each player best-responds to the other's agreed action rather than to the equilibrium distribution of the other's actions. In the limit where the mistake probability vanishes, this does not matter, however.

Whichever target they set, they should choose the agreement with the smallest marginal incentives to breach. To see this notice first that the payoff-difference between choosing m_i and choosing $m_i - 1$ is given by $\beta_i(m, \theta_i) = \gamma_i(m) - \theta_i \eta_i(m)$. Thus $ln(q(m_i|m)) - ln(q(m_i - 1|m)) = (1/\mu)\beta_i(m, \theta_i)$ and the log ratio of the probability of keeping the promise to the probability of marginally deviating increases whenever the marginal incentive to breach decreases. Thus the probability of keeping the promise is larger than the probability of breach if and only if the marginal incentive to breach is negative. If $\beta_i(m, \theta_i)$ is indeed negative, then choosing any strategy $s_i < m_i - 1$ incurs even lower payoff than the marginal breach due to the concavity of underlying payoffs (see the proof of Proposition 1). Parties who seek to minimize the probability of underprovision must thus seek to minimize the marginal incentives to breach. In the limit where $\mu \to 0$, the players are almost sure to choose the optimal strategy. Yet, the probability of keeping one's promise still depends on the agreement and it is maximized precisely where the marginal incentive to breach is minimized.[28]

The probability of breaching the agreement thus decreases if and only if the marginal incentive to breach decreases. By Proposition 3 then, when actions are strategic substitutes, the underlying game equilibrium has better safeguards against breaching than any agreement that constitutes a Pareto-improvement to the equilibrium. By Lemma 1 and Theorem 1, symmetric increments of agreed actions decrease the marginal incentives to breach if $\phi_i + \theta_i \phi_j > -\delta_i - \theta_i \sigma_i$ for $i = 1, 2$ and they increase the marginal incentives to breach if the opposite strict inequality holds true. Thus, when actions are sufficiently strong strategic complements, negotiators interested in maximizing the probability that the agreement is kept will implement all symmetric increments to the agreement. Once further increments are bounded by the maximal action of either player, the players will agree to increase or decrease the other player's action if and only if this decreases the joint probability of breach (17) or the larger of breach probabilities (18). (Recall that the marginal incentive to breach of i increases and that of j decreases in m_i, so the breach probabilities move to opposite directions when only one agreed action is changed.) So an agreement where at least one player's agreed action is her maximal action will be selected. In symmetric games with symmetric players each player will promise to choose the maximal action. In symmetric games with weaker strategic complements or with strategic substitutes, $\phi + \theta\phi < -\delta - \theta\sigma$, each player will promise to choose the interior UG equilibrium action.[29] This

[28]The payoff of choosing an action $s_i > m_i$ is independent of which of the actions smaller than s_i the player has actually agreed to choose and if $s_i > max\{BR_i(m_j)\}$, then the payoff is smaller than the payoff of keeping one's promise. Thus $ln(q(s_i|m)) - ln(q(m_i|m)) < 0$ and

$$lim_{\mu \to 0} \frac{q_i(m_i|m)}{q_i(s_i|m)} = \infty.$$

[29]Recall that profiles where each player chooses an action smaller than the equilibrium action are never incentive compatible, nor are profiles where there is a player, say i, who chooses an action lower than the UG best reply to m_j. This proves the first part of the result. When $\theta_1 = \theta_2$ and when the game is symmetric, the interior

yields the following proposition:

Proposition 4 *If $\phi_i + \theta_i\phi_j > -\delta_i - \theta_i\sigma_i$ for $i = 1, 2$, then*

$$lim_{\mu\to 0}\tilde{m}(\mu), lim_{\mu\to 0}\hat{m}(\mu) \in \{s|\exists i : s_i = n_i\}$$

If $\phi_i + \theta_i\phi_j < -\delta_i - \theta_i\sigma_i$ for $i = 1, 2$, then

$$lim_{\mu\to 0}\tilde{m}(\mu), lim_{\mu\to 0}\hat{m}(\mu) \in \{s|\exists i : s_i = s_i^*\}.$$

In particular, this is the case when $\phi_i < 0$ for $i = 1, 2$.

In symmetric games with $\theta_i = \theta_j = \theta$, if $\phi + \theta\phi > -\delta - \theta\sigma$, then

$$lim_{\mu\to 0}\tilde{m}(\mu) = (n_1, n_2) = lim_{\mu\to 0}\hat{m}(\mu).$$

If $\phi + \theta\phi < -\delta - \theta\sigma$, then

$$lim_{\mu\to 0}\tilde{m}(\mu) = s^* = lim_{\mu\to 0}\hat{m}(\mu).$$

Minimizing the risk of breach may of course not be the only objective of the negotiators. They might also wish to maximize the surplus generated by the agreement (if side-payments are allowed for) or to ensure Pareto-efficiency (when side-payments are not feasible). When $\delta_i + 2\phi_i + \sigma_i \geq 0$ for $i = 1, 2$, the Pareto-efficient agreements lie in $\{s|\exists i : s_i = n_i\}$ - the set where the agreement most likely to be kept also lies - and one of these agreements is also surplus maximizing. This is yet another way to illustrate that efficiency and incentives are aligned with sufficiently strong strategic complements. When $\delta_i + 2\phi_i + \sigma_i < 0$, the Pareto-efficient and surplus-maximizing agreements tend to lie in the interior rather than on the boundary of the strategy sets (this of course depends on the stringency of the bounds of each strategy set) and the players simply minimizing the risk of breach tend to compromise on surplus and choose too high contributions if $\phi_i + \theta_i\phi_j > -\delta_i - \theta_i\sigma_i$ for $i = 1, 2$ or too small contributions if $\phi_i + \theta_i\phi_j < -\delta_i - \theta_i\sigma_i$ for $i = 1, 2$.

I will formalize the negotiators' trading off between the risk of breach and the generated surplus in my second set-up as follows. Suppose that there is uncertainty about the proneness to guilt types of the two players. This uncertainty will be fully resolved between getting to an agreement and choosing the actions. That is, the negotiators will choose an agreement behind a veil of ignorance, but when the players choose actions given an agreement, the veil will have have been lifted and proneness to guilt types will be common knowledge. Given the assumption that there is no guilt about breaching if the opponent breaches too, the players of a game with strategic substitutes will abide by the deal if and only if each proneness to guilt type satisfies (11) and they coordinate on the underlying game equilibrium s^* otherwise. Let us denote the

equilibrium must be symmetric and the breaching incentives at symmetric profiles must be equal. Thus the latter part of the result follows.

corresponding threshold type of player i by $\hat{\theta}_i(m)$ and define $\hat{\theta}_{max}(m) = max\{\hat{\theta}_1(m), \hat{\theta}_2(m)\}$. If the distributions of the proneness to guilt types are symmetric and have a cumulative distribution function $F(\theta_i)$, then negotiators maximizing the expected surplus will agree on the agreement which satisfies

$$max_m\{(u_1(m) + u_2(m))(1 - F(\hat{\theta}_{max}(m))) + (u_1(s^*) + u_2(s^*))F(\hat{\theta}_{max}(m))\} \tag{19}$$

in a game with strategic substitutes. Let us consider symmetric increments: $m_1 - s_2^* = m_1 - s_2^*$. With strategic substitutes, as the surplus is increased the probability of breach also increases. Therefore the maximization of surplus conditional on compliance to the agreement must be compromised due to the risk of breach. To the contrary when $\phi_i + \theta_i\phi_j > -\delta_i - \theta_i\sigma_i$ for all θ_i in the support of the distribution and moreover $\delta_i + 2\phi_i + \sigma_i \geq 0$ for $i = 1, 2$ (strategic complementarity is a necessary condition for all inequalities), then the negotiators maximizing expected surplus will implement every symmetric change to the agreement. They will thus agree on an ex-post Pareto-efficient profile. In this case the maximization of surplus conditional on compliance must be compromised only to the extent of selecting between the Pareto-efficient agreements. In symmetric games, no such compromises are needed, and the negotiators will select a surplus-maximizing agreement which is also the agreement which minimizes the risk of breach. When actions are weaker strategic complements so that $\delta_i + 2\phi_i + \sigma_i < 0$ for $i = 1, 2$ but nevertheless $\phi_i + \theta_i\phi_j > -\delta_i - \theta_i\sigma_i$ for all θ_i in the support of the distribution, the players might agree on higher than Pareto-efficient actions when trading-off ex-post surplus with higher probability of compliance.

5 Cournot duopoly

In this section I study a linear cournot duopoly with heterogeneous goods. The purpose is twofold: first, to provide an example how the model can be applied; second, to illustrate how the model can be accomodated to an infinite game framework with twice continously differentiable payoffs (see footnote 13).

There are two firms indexed by $i, j = 1, 2$ where $i \neq j$. The profits of firm i are $\pi_i(q_i, q_j) = (1 + \phi q_j - \frac{1}{2}q_i)q_i$. Each firm chooses its supply from an interval $q_i \in [0, n]$ where $n > 1$. Parameter ϕ reflects the product complementarity between the final goods produced by the two firms. The price for firm i's good, $(1 + \phi q_j - \frac{1}{2}q_i)$, increases in j's supplied quantity if and only if $\phi > 0$ indicating that the goods are complements. To the contrary when $\phi < 0$ the goods are substitutes. Notice moreover that the commonly studied case of homogenous goods corresponds to $\phi = -1/2$. Product complementarity is distinct but closely related to strategic complementarity as we shall see shortly. Firm i's marginal payoff (marginal private return) is

$$\frac{\partial \pi_i}{\partial q_i} = 1 - q_i + \phi q_j$$

and the marginal effect of i's produced quantity on j's profit (marginal external return) is

$$\frac{\partial \pi_j}{\partial q_i} = \phi q_j. \tag{20}$$

The payoff increases strictly in the opponent's action (the condition imposed in Section 4) if and only if $\phi > 0$. Yet as pointed out in Section 4, games where the payoff strictly decreases in the opponent's action, such as the cournot duopoly with substitute products $\phi < 0$, can be analyzed if one first reverses the ordering of each strategy set and uses the transformation $\tilde{q}_i = -q_i$ where $\tilde{q}_i = [-n_i, 0]$. The payoff π_i of player i can now equivalently be written as $\tilde{\pi}_i(\tilde{q}_i, \tilde{q}_j) = -(1 - \phi\tilde{q}_j + \frac{1}{2}\tilde{q}_i)\tilde{q}_i$, yielding

$$\frac{\partial \tilde{\pi}_j}{\partial \tilde{q}_i} = \phi \tilde{q}_j,$$

which is positive[30] as required in Section 4.

From now on, the transformation is applied to the action sets and payoffs if and only if the goods are substitutes. Whether the goods are substitutes or complements, the private and external returns are weakly decreasing, $\frac{\partial^2 \pi_i}{\partial q_i^2} = -1 = \frac{\partial^2 \tilde{\pi}_i}{\partial \tilde{q}_i^2}$, $\frac{\partial^2 \pi_j}{\partial q_i^2} = 0 = \frac{\partial^2 \tilde{\pi}_j}{\partial \tilde{q}_i^2}$, as assumed throughout this paper. Similarly, independently of whether the transformation is carried out, the marginal effect of j's supply on the marginal private and external return of i's supply[31] is given by

$$\frac{\partial^2 \pi_i}{\partial q_i \partial q_j} = \phi = \frac{\partial^2 \tilde{\pi}_i}{\partial \tilde{q}_i \partial \tilde{q}_j},$$

implying that goods are complementary if and only if the supply choices are strategic complements. In summary, the transformation makes payoffs increasing in the opponent's action and yet it preserves symmetry, strategic complementarity, and decreasing marginal private and external returns.

Let's us now focus on the case of substitute products and consider the transformed action sets and payoffs. Since $\delta_i = -1 < 0$, the first order condition characterizes the best reply curve[32] when $1/\phi_i < \tilde{q}_j < (1 - n_i)/\phi_i$:

$$BR_i(\tilde{q}_j) = \begin{cases} -n_i & if \quad \tilde{q}_j \geq \frac{n_i - 1}{\phi} \\ 0 & if \quad \tilde{q}_j \leq \frac{1}{\phi} \\ -1 + \phi\tilde{q}_j & otherwise \end{cases} \tag{21}$$

In the relevant range, the slope of the best-reply curve is $\phi < 0$. The unique Nash equilibrium lies at the intersection of the best-reply curves at $\tilde{q}_i^* = -1/(1 - \phi)$. The joint-profit-maximizing

[30] Both ϕ and $\tilde{q}_j$ are negative.

[31] Any two-player submodular game, which in our setting is equivalent with $\phi < 0$, can be transformed into an equivalent game which is supermodular by setting $\widetilde{q}_2 = -q_2$. However, then both payoffs are not increasing in the action of the opponent. See Vives (2005, pp.451) for instance.

[32] Best-reply curves are straight lines and best-responds are unique.

actions are $\tilde{q}_i = -1/(1-2\phi) = \tilde{q}_j$. Each player's action is greater at the surplus maximizing profile than at the equilibrium given that ϕ is negative, as generally noted in Section 4.[33] Condition (7) requires that $i's$ marginal payoff is non-positive, so that a deviation does not make both better off. In the transformed Cournot-game, this translates to a requirement that the agreement must lie within the following set $m \in \{\tilde{q}| -1 - \tilde{q}_i + \phi\tilde{q}_j \leq 0\}$. An agreement which is incentive compatible for both must lie to the north-east of both the downward-sloping best-reply curves where each firm is tempted to choose a lower $\tilde{q}_i$ than agreed on (corresponding to a tendency to produce more than agreed on in the equivalent pre-transformation game). Proposition 1 states that within this set a non-positive marginal incentive to breach is necessary and sufficient for incentive compatibility. The marginal incentive to breach reads

$$1 + \tilde{q}_i - \phi\tilde{q}_j - \theta_i\phi\tilde{q}_j. \tag{22}$$

This expression evinces the incentive effects of increasing each action. A greater action of player i raises i's marginal gain and leaves the marginal harm unaffected and consequently the marginal incentive to breach is increasing in i's own action. Since the game has strategic substitutes, $\phi < 0$, a greater action of player j lowers the marginal harm on j. A higher action of j also raises i's marginal gain. Accordingly, the incentive to breach increases as partners strike an agreement to raise joint profits as generally shown in Section 4.

Let us derive an upper bound for a credible promise by i. Let us denote the agreement by $m = \tilde{q}^* + k = (m_1 = \tilde{q}_1^* + k_1, m_2 = \tilde{q}_2^* + k_2)$. Setting (22) less or equal to zero and solving in terms of k_i yields

$$k_i \leq -1 + \phi(\tilde{q}_j^* + k_j)(1 + \theta_i) - \tilde{q}_i^*. \tag{23}$$

This upper bound is declining in k_j since ϕ is negative. Substituting from the equilibrium condition $\phi\tilde{q}_j^* = 1 + \tilde{q}_i^*$ to the right hand side yields $k_i \leq (1 + \tilde{q}_i^*)(1 + \theta_i) + \phi k_j(1 + \theta_i) - \tilde{q}_i^*$ which is increasing in ϕ: in the linear duopoly model k_i can be pushed further away from the equilibrium as strategic substitutability becomes stronger.

Let us now analyze informal agreements between firms that produce complementary goods. The price for i's product is higher as j sells more and accordingly i's payoff increases in j's production. Therefore no transformation of the payoff function is needed to satisfy the assumptions of Section 4. The slope of each best-reply curve is $\phi > 0$. The Nash equilibrium quantity equals $1/(1-\phi)$ when $\phi < (n_i - 1)/n_i$. The latter is also the condition for joint-profit-maximizing supply being greater than the equilibrium quantity, and there is room for Pareto-improvements through pre-play agreements. The marginal incentive compatibility condition for i,

$$-1 + q_i - \phi q_j - \theta_i\phi q_j \leq 0, \tag{24}$$

expresses that the marginal gain is decreasing and the harm is increasing in the opponent's produced quantity as shown in Section 4 for games with strategic complements. The upper

[33] This greater strategy is associated with a lower than equilibrium quantity, $1/(1-2\phi) < 1/(1-\phi) = q_i^*$ when the non-transformed game is used.

bound on the credible promise of i is $k_i \leq 1+\phi(q_j^*+k_j)(1+\theta_i)-q_i^*$. When $1-\phi(1+\theta_i)>0$, we can explicitly solve for an upper bound for how much production can be symmetrically increased above the equilibrium

$$k \leq min_i\{\frac{-1+q_i^*+\phi q_j^*+\theta_i\phi q_j^*}{1-\phi(1+\theta_i)}\}, \tag{25}$$

where $-1+q_i^*+\phi q_j^*=0$ since q^* is an underlying game equilibrium. However, if

$$1-\phi(1+\theta_i)<0 \tag{26}$$

for $i=1,2$, the players can agree on any symmetric production profile where firms produce more than in the equilibrium - including the quantity which maximizes the joint profits. When inequality (26) holds for $i=1,2$, and $\phi<(n_i-1)/n_i$, the unique underlying game equilibrium is inefficient and an informal agreement on the joint-profit-maximizing profile substantially raises the firms' profits.

6 Discussion

This paper proposes a simple model of informal pre-play agreements. It studies the interaction between intrinsic motivation to abide by agreements and the incentive structure of the underlying game. In the model, the incentives to comply with an agreement are generally sustained by intrinsic motivation the effectiveness of which depends on the strategic details of the underlying game.[34] Breaching an agreement is assumed to induce a guilt cost. The cost is postulated to increase in the harm that breaching inflicts on the opponent. This presupposition aligns with the findings of Gneezy (2005) and Sutter (2009) and a host of literature in other social sciences (Hoffman, 1982; Hamilton and Rytina, 1980; Darley et al., 2000; Carlsmith et al., 2002). The model illustrates that the effectiveness of guilt in aligning the parties' incentives and in promoting their welfare depends on the strategic nature of the underlying interaction: in partnerships where inputs are strategic substitutes, the Pareto-efficiency of an agreement is in conflict with both the extrinsic and intrinsic incentives to abide by the agreement: marginally breaching an agreement on a Pareto-improvement to an underlying game equilibrium is associated with greater private gain and a smaller harm on the opponent than marginally breaching an agreement to play the underlying game equilibrium itself. Nevertheless, partnerships with sufficiently strong strategic complements may avoid this conflict. Moreover, a Pareto-efficient agreement is incentive compatible. This holds even when strategic complementarity is so weak that the unique underlying game equilibrium is inefficient. I also show that if parties negotiating an agreement wish to maximize the probability that the agreement is kept (when they make mistakes in the implementation of their actions) or maximize the expected surplus (when there is uncertainty about the proneness to guilt types), they tend to choose a Pareto-efficient agreement if actions

[34] To my knowledge, there are fairly few models of informal agreements. Yet, see recent independent contributions by Kessler and Leider (2012) and Dufwenberg et al. (2012).

are sufficiently strong strategic complements but they will compromise on efficiency if the actions are strategic substitutes.

A related exercise would consider negotiators who choose agreements according to fairness ideals and ask how well the chosen agreement fares in terms of the incentives to abide by the agreement. An interesting related avenue for further comparative study of social norms is proposed by Kranz (2010). He suggests that the prevalent social norm might be selected in order to maximize the utility of compliant types as a function of the context and the population distribution of privately known proneness to guilt types (complier-optimal norms). Several interesting questions arise when casting his approach over the present one and allowing guilt to depend on the harm inflicted on compliant others: for instance, whether the complier-optimal norms would tend to select first-best efficient norms when actions are strategic complements.

Apart from Section 4.2, I have held the assumption that proneness to guilt types are complete information and parties are rational or close to rational in the sense of being highly likely to choose their optimal strategies given their commonly known incentives. These assumptions are surely not entirely plausible. For instance there is no reason to expect that experimental subjects would know each other's disposition to guilt. Will the present model be able to capture the comparative statics patterns of strong effects of communication under strategic complements and weak effects under substitutes if one relaxes the complete information and rationality assumptions? Results on monotone comparative statics (Vives, 2005; Vives and van Zandt, 2007; Schmutzler 2011) suggest that this should be the case. In particular, each player's payoff function satisfies weak increasing differences in proneness to guilt and actions are strategic complements. Moreover, under incomplete information about proneness to guilt, types are affiliated in the sense that without an agreement each player disregards, and is known to disregard, the externality on the other while with an agreement in place both have a weakly positive concern for the payoff of the other.[35] Therefore each player expects higher actions by the opponent when an agreement is in place than when not. All these factors imply that agreements shift actions upwards in the sense of first-order stochastic dominance.[36] Schmutzler's findings suggest that the comparative statics predictions would continue to hold for solution concepts incorporating features of noise and bounded rationality, such as the quantal-response equilibrium or some belief-learning dynamics, concepts which are better suited for descriptive work.

[35]This trivially implies log-supermodularity of the type distribution, as required in Vives and van Zandt (2007), when comparing conditions with and without an agreement.

[36]This result does not generally hold if actions are substitutes: although the assumption of increasing differences in the proneness to guilt parameter is satisfied and thus each player's incentives to increase her action are individually strengthened, increasing a players proneness to guilt strengthens that player's incentive to increase her action and thereby, due to strategic substitutability, creates an incentive for the opponent to lower his action.

7 Appendix - The Proofs

The starred equations and theorems refer to the main article. For instance (6*) refers to Equation (6) in the main article.

To simplify exposition, we adopt the following concepts. For $m \in S$ and for $k \in \mathbb{Z}$, let us call $m + k = (m_1 + k, m_2 + k)$ a symmetric change of actions by k vis-à-vis m. For $k \in \mathbb{Z}$, the effect on the marginal gain, on the marginal harm, and on the agreed payoff due to such a change are thus $\gamma_i(m + k) = \gamma_i(m_i + k,\ m_j + k)$, $\eta_i(m + k) = \eta_i(m_i + k, m_j + k)$ and $\pi_i(m+k) = \pi_i(m_i+k, m_j+k)$, respectively. The proofs are given applying the general functional form of guilt cost, $\theta_i max\{h_j(m, s_i), 0\}^{\varphi}$. Thus the the marginal incentive to breach can be defined as $\beta_i(m, \theta_i) = \gamma_i(m) - \theta_i max\{\eta_j(m), 0\}^{\varphi}$.

7.1 Proof of Proposition 1

Proof. We will show that there is a deviation that pays off if and only if $\beta_i(m, \theta_i) > 0$. Let $\beta_i(m, \theta_i) > 0$. By the definition of $\beta_i(m, \theta_i), \gamma_i(m) - \theta_i max\{\eta_j(m), 0\}^{\varphi} > 0$ and there is a deviation which pays off.

Let there be a deviation which pays off. Suppose to the contrary that $\beta_i(m, \theta_i) \leq 0$ and thus

$$\pi_i(m_i - 1, m_j) - \pi_i(m_i, m_j) \leq \theta_i max\{h_j(m, m_i - 1), 0\}^{\varphi}. \tag{27}$$

There are two cases to consider $\pi_i(m_i - 1, m_j) - \pi_i(m_i, m_j) < 0$ and $\pi_i(m_i - 1, m_j) - \pi_i(m_i, m_j) \geq 0$. In the first case, it is also true that $\pi_i(m_i, m_j) - \pi_i(m_i + 1, m_j) < 0$ since otherwise m_i is an underlying best-reply to m_j by the concavity of π_i in its first argument. Now $\pi_i(m_i, m_j) - \pi_i(m_i + 1, m_j) < 0$ implies that $m_i \notin M_i^F$ which is a contradiction. In the second subcase $\pi_i(m_i - 1, m_j) - \pi_i(m_i, m_j) = \gamma_i(m) > 0$ and thus $h_j(m, m_i - 1) > 0$ since $\beta_i(m, \theta_i) \leq 0$. By assumption, harm increases in deviations further downwards. Also by assumption guilt cost is convex in h_j and π_j is concave in s_i. Thus harm is convex in s_i and the guilt cost is also convex in s_i as a composite of two convex functions. On the other hand by assumption, the payoff π_i is concave in s_i and thus the gain from breaching $\pi_i(s_i, m_j) - \pi_i(m_i, m_j)$ is concave in s_i. Thus if $\beta_i(m, \theta_i) \leq 0$ then no deviation pays off. We have a contradiction. ■

7.2 Proof of Lemma 1*

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

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

$$\eta_j(m_j, m_i + 1) - \eta_j(m_j, m_i)$$

$$= \pi_j(m_j, m_i + 1) - \pi_j(m_j, m_i) - [\pi_j(m_j, m_i) - \pi_j(m_j, m_i - 1)]$$
$$= \sigma_j$$

$$\eta_j(m_j + 1, m_i) - \eta_j(m_j, m_i)$$
$$= \pi_j(m_j + 1, m_i) - \pi_j(m_j + 1, m_i - 1) - [\pi_j(m_j, m_i) - \pi_j(m_j, m_i - 1)]$$
$$= \phi_j \blacksquare$$

7.3 Lemma 2

Lemma 2 *Let $\varphi \geq 0$ so that guilt cost is convex in the inflicted harm. Let $-\sigma_j \leq \phi_j$ Let there exist a k satisfying*

$$\gamma_i(s^*) - (\delta_i + \phi_i)k \leq \theta_i max\{\eta_j(s^*) + (\sigma_j + \phi_j)k, 0\}^{\varphi}$$

where $max\{\eta_j(s^*) + (\sigma_j + \phi_j)k, 0\}^{\varphi}$ must be positive whenever $\gamma_i(s^*) - (\delta_i + \phi_i)k$ is positive. In that case

$$\theta_i \geq \frac{-(\delta_i + \phi_i)}{\{\eta_j(s^*) + (\sigma_j + \phi_j)k\}^{\varphi} - \{\eta_j(s^*) + (\sigma_j + \phi_j)(k-1)\}^{\varphi}} \geq 0.$$

Then every $s^* + k'$ with $k' > k$ is incentive compatible for i.

Proof. By the condition in Proposition 1*, an agreement is incentive compatible for i iff

$$\gamma_i(m) \leq \theta_i max\{\eta_j(m), 0\}^{\varphi},$$

where $\eta_j(m) > 0$ since payoffs are increasing in the opponent's action. The fundamental theorem of calculus allows writing this as

$$\gamma_i(s^*) - (\delta_i + \phi_i)k \leq \theta_i max\{\eta_j(s^*) + (\sigma_j + \phi_j)k, 0\}^{\varphi} \tag{28}$$

where I have used Lemma 1. Thus the when the first condition in Lemma 2 holds, the agreement $s^* + k$ is incentive compatible for i. When moreover

$$\theta_i \geq \frac{-(\delta_i + \phi_i)}{\{\eta_j(s^*) + (\sigma_j + \phi_j)k\}^{\varphi} - \{\eta_j(s^*) + (\sigma_j + \phi_j)(k-1)\}^{\varphi}} \geq 0$$

(where one should remark that $\sigma_j + \phi_j) \geq 0$), the right hand side of (7.3) rises faster than the left hand side as a function of k. ■

7.4 Proof of Theorem 2*

Proof.

Let us first assume that $\phi_i > -\delta_i$.

Let us first look for the best action profile for each player that can be reached through symmetric changes of both actions $argmax_k\{\pi_i(s^* + k)\}$.

Since s^*+k Pareto-dominates s^*, it must be that $k>0$ by the same argument as in the proof of Proposition 3*. Thus for a symmetric change of actions to be Pareto-improving, it must hold that $k>0$.

There may be no symmetric change of actions that would reach a Pareto-efficient profile. This is the case if at $s^*+\overline{k}$ as defined above $\gamma_j(s_j^*+\overline{k}+1, s_i^*+\overline{k})<0$, in which case j can improve the payoff of both by deviating upwards. Let thus j's action be increased until her UG best-response is reached (which can satisfy $s_j=n_j$). Since j's agreed action is unilaterally increased, by Lemma 1*, γ_i is weakly smaller.

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

Let know (14*) hold, since s^*+k Pareto-dominates s^*, it must be that $k>0$ by the same argument as in the proof of Proposition 3*. Thus for a symmetric change of actions to be Pareto-improving, it must hold that $k>0$. Lemma 2 ensures that for every $K>k$, s^*+K is incentive compatible for i.

Let us show that i's payoff-maximizing profile among symmetric increases of actions, denote it by $s^*+\overline{k}_i$, exists and satisfies $\overline{k}_i \geq k$. To see this, suppose first that $\sigma_i+\delta_i+2\phi_i \geq 0$. The assumption implies that $\pi_i(s+K)$ is convex in K. Now, by the argument above, $\pi_i(s^*)-\pi_i(s^*-K)>0$ for any $K>0$. Therefore, every symmetric increase of actions increases the payoff of i vis a vis s^* and by the boundedness of the strategy space, there exists $\overline{k}_i$ such that $s^*+\overline{k}_i$ maximises i's payoff. Suppose alternatively that $\sigma_i+\delta_i+2\phi_i<0$. Then $\pi_i(s+K)$ is strictly concave in K. By assumption, $\pi_i(s^*+k)\geq \pi_i(s^*+k-1)$. Since the strategy set is finite, a maximizer $s^*+\overline{k}_i$ among symmetric increases of actions exists and it satisfies $\overline{k}_i \geq k$. By the same arguments, there is a $\overline{k}_j \geq k$ which maximizes j's payoff among the symmetric changes of actions. Let $\overline{k}\equiv min(\overline{k}_i, \overline{k}_j)$. Notice that by the argument above, $s^*+\overline{k}$ is incentive compatible for i. Let us show that $s^*+\overline{k}$ is Pareto-efficient if $\phi_j+\delta_j<0$. (We will return further below to the case $\phi_j+\delta_j\geq 0$.) It is easy to see that no symmetric change of actions is Pareto-preferred to $s^*+\overline{k}$. Moreover no unilateral change of an action is Pareto-preferred since s^* is an equilibrium and thus, by Lemma 1*, $\gamma_i(s_i^*+K+1, s_j^*+K)\geq 0$ for $K\geq 0$ and payoffs are concave in own action. Thus i's payoff is worse if her action alone is increased from $s^*+\overline{k}$. On the other hand, if i's action is decreased, j's payoff decreases by assumption. Combinations of symmetric changes of actions and unilateral changes of one particular action reach any strategy where $s_i>s_i^*$ yielding the result.

If $\phi_j+\delta_j\geq 0$, there may be no symmetric change of actions that is Pareto-efficient. This is the case if at $s^*+\overline{k}$ as defined above $\gamma_j(s_j^*+\overline{k}+1, s_i^*+\overline{k})<0$, in which case j can improve the payoff of both by deviating upwards. Let thus j's action be increased until her UG best-

response is reached (which can satisfy $s_j = n_j$). Since j's agreed action is unilaterally increased, by Lemma 1*, γ_i is weakly smaller and η_j is weakly greater than before the unilateral increase of j's action. Thus, i's marginal incentive to breach is smaller or equal to that at $s^* + \overline{k}$.

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

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

Acknowledgements

I am indebted to Steffen Huck and Philippe Jehiel for continous support, encouragement, and insightful comments. Further I would like to thank Martin Dufwenberg, Tore Ellingsen, Daniel Friedman, Antonio Guarino, Werner Güth, Fabio Michelucci, David Myatt, Birendra Kumar Rai, Randolph Sloof, Joel Sobel, Rune Stenbacka, Christoph Vanberg, Juuso Välimäki, Jörgen Weibull, Georg Weizsäcker, and many seminar audiences. I also gratefully acknowledge the Financial support of the Yrjö Jahnsson Foundation.

References

[1] Anderson, S. P., Goeree, J. K., and Holt C. A., 1998. A theoretical analysis of altruism and decision error in public goods games. J. Public Econ. 70, 297-323.

[2] Battigalli, P., Dufwenberg, M., 2007. Guilt in Games. Amer. Econ. Rev., Papers & Proceedings 97, 170-176.

[3] Battigalli, P., Dufwenberg, M., 2009. Dynamic Psychological Games. J. Econ. Theory 144, 1-35.

[4] Bicchieri, C., 2006. The Grammar of Society. New York: Cambridge University Press.

[5] Carlsmith, K.M., Darley, J.M, Robinson, P.H., 2002. Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. J. Pers. Soc. Psychol. 83, 284-299.

[6] Charness, G., Dufwenberg, M., 2006. Promises and Partnership. Econometrica 74, 1579-1601.

[7] Charness, G., Dufwenberg, M., 2011. Participation. Amer. Econ. Rev. 101, 1213-39.

[8] Crawford, V., P., 2003. Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions. Amer. Econ. Rev. 93, 133-149.

[9] Crawford, V., P., Sobel, J., 1982. Strategic Information Transmission. Econometrica 50, 1431-51.

[10] Darley, J. M., Carlsmith, K. M., Robinson, P. H., 2000. Incapacitation and just deserts as motives for punishment. Law Hum. Behav. 24, 659-684.

[11] Demichelis, S., Weibull, J., 2008. Meaning and Games: A Model of Communication, Coordination and Games. Amer. Econ. Rev. 98, 1292-1311.

[12] Dufwenberg, M., Servatka, M., Vadovic, R. (2012). ABC on Deals. Manuscript, University of Arizona.

[13] Ellingsen, T. , Johannesson, M., 2004. Promises, Threats, and Fairness. Econ. J. 114, 397-420.

[14] Ellingsen, T, Ostling, R., 2010. When Does Communication Improve Coordination? Amer. Econ. Rev. 100, 1695-1724.

[15] Erat, S, Gneezy, U., 2011. White Lies. Management Science. In press.

[16] Farrell. J., 1987. Cheap Talk, Coordination, and Entry. Rand J. Econ. 18, 34-39.

[17] Farrell, J., 1988. Communication, Coordination and Nash Equilibrium. Econ. Lett. 27, 209-214.

[18] Farrell, J., Rabin M., 1996. Cheap Talk. J. Econ. Perspectives 10, 103-118.

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

[20] Geanokoplos, J., Pearce, D., Stachetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60-79.

[21] Gneezy, U., 2005. Deception: The Role of Consequences. Amer. Econ. Rev. 95, 384-394.

[22] Hamilton, V.L., Rytina, S., 1980. Social Consensus on Norms of Justice: Should Punishment Fit the Crime? Amer. J. Sociol. 85, 1117-1144.

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

[24] Isaac, M., McCue, K., Plott C., 1985. Public Goods Provision in an Experimental Environment. J. Public Econ. 26, 51-74.

[25] Isaac, M., Walker J., 1988. Communication and Free-riding Behavior: the Voluntary Contribution Mechanism. Econ. Inq. 26, 586-608.

[26] Kandel, E., Lazear, E., 1992. Peer Pressure and Partnerships. J. Polit. Econ. 100, 801-817.

[27] Kaplow, L., Shavell, S., 2007. Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System. J. Polit. Econ. 115, 494-514.

[28] Kessler J.B., Leider, S. (2012) Norms and Contracting. Manage. Sci. 58, 6277.

[29] Kranz, S., 2010. Moral Norms in a Partly Compliant Society. Games Econ. Behav. 68, 255-274.

[30] Ledyard, J.,O., 1995. Public Goods: A Survey of Experimental Research. In: Kagel, J., Roth, A. E. (Eds) Handbook of Experimental Economics. Princeton, NJ: Princeton University Press.

[31] Lopez-Perez, R., 2008. Aversion to Norm-Breaking: A Model. Games Econ. Behav. 64, 237-267.

[32] Lopez-Perez, R., 2012. The Power of Words: A Model of Honesty and Fairness. J. Econ. Psychol., 33, 642658.

[33] Miettinen, T., Suetens, S., 2008. Communication and Guilt in a Prisoner's Dilemma. J. Confl. Resolut. 52, 945-960.

[34] Milgrom, P., Roberts, J., 1990. Rationalizability, Learning and Equilibrium in Games with Strategic Complementarities. Econometrica 58, 1255-1277.

[35] Sally, D., 1995. Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992. Ration. Soc. 7, 58-92.

[36] Schmutzler, A. 2011. A Unified Approach to Comparative-Statics Puzzles in Experiments, Games Econ. Behav. 71, 212-223.

[37] Sebald, A. 2010. Attribution and Reciprocity. Games Econ. Behav. 68, 339352.

[38] Suetens, S., 2005. Cooperative and Noncooperative R&D in Experimental Duopoly Markets. Int. J. of Ind. Organ. 23, 63-82.

[39] Sutter, M., 2009. Deception through telling the truth?! Experimental evidence from individuals and teams. Econ. J. 119, 47-60.

[40] Topkis, D., 1979. Equilibrium Points in Non-Zero Sum n-Person Sub-modular Games. SIAM J. Control and Optim., 17, 773787.

[41] Vanberg, C., 2008. Why Do People Keep Promises? An Experimental Test of Two Explanations. Econometrica 76, 1467 - 1480.

[42] Vives, X. 2005. Complementarities and Games: New Developments. J. Econ. Lit. 19, 305 - 321.

[43] Vives, X. van Zandt, T. 2007. Monotone equilibria in Bayesian games of Strategic Complementarities. J. Econ. Theory 134, 339-360.