Agreeing on Play with a Dislike to Lie - A Theory of Preplay Negotiation*

Topi Miettinen†

June 28, 2005

Abstract

Experimental evidence suggests that communication increases cooperation in the prisoner's dilemma and contributions in public good games. This paper claims that this efficiency effect may be driven by the guilt that is felt about breaching informal agreements.

Informed by findings in social psychology, we construct a two-player model of preplay negotiation with players who are prone to feelings of guilt. We confirm the efficiency effect in the prisoner's dilemma and in the public good game and derive many other testable predictions for these games.

We call a strategy profile agreeable if agreements to play according to them would not be broken and if both players have an incentive to reach such an agreement. In supermodular games where payoff increases in opponent's action, if any non-equilibrium action profile above the equlibrium is agreeable, then an outcome in the core of the game is agreeable. Symmetric submodular games, for instance, fail such a property.

JEL Classification C72, C78, Z13

KEYWORDS: preplay negotiation, guilt

There is no commonly honest man who does not more dread the inward disgrace of such an action, the indelible stain which it would for ever stamp upon his own mind, that the greatest external calamity which, without any fault of his own, could possibly befal him; and 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))

*Preliminary, do not use or cite without permission. Financial support of the Yrjö Jahnsson Foundation gratefully acknowledged. I am very grateful to Steffen Huck and Philippe Jehiel for their advice, encouragement, suggestions, and discussions. Also I would like to thank Mikko Leppämäki, Francesco Squintani, and Pekka Sääskilahti and the seminar participants in Copenhagen (CEBR), Helsinki (HECER), and London (UCL) for comments. All errors are mine.

†Deparment of Economics, University College London, Gower Street WC1E 6BT, t.miettinen@ucl.ac.uk.

1 Introduction

Ray and Cal have a magic pot and ten dollar coins each. Each coin 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. Cal figures that, whatever Ray puts into the pot, for each dollar he puts into the pot, he gets only $\frac{3}{4}$ dollars back and, hence, should put no coins into the pot.

Before they decide, they can talk to each other. They may agree on how many coins each of them puts into the pot. The agreement is not binding. Yet, having talked to Ray for a while, he seems like nice guy to Cal. And Cal starts to think that he would feel bad if he breached an agreement. He also figures that Ray may well think similarly about him. Eventually, Ray and Cal agree on putting two 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 thus far he has seen few models where people care about causing harm to the other by not doing as agreed.

This paper presupposes that both the magic pots and the dislike to breach oral agreements are worth taking seriously. The *agreement (an agreed action)* is an action profile (an action) of the *underlying game,* the game that is played after the negotiations. Having agreed on a profile, and if the opponent does her part of the deal, a player who breaches may feel guilt which lowers her utility, but only if the action profile dominates the underlying game equilibrium payoff. Thereby, pareto improvements may be reached through communication.

We let the underlying game be any normal form game. Given such a game and players' proneness to guilt, each agreement maps the game into another game with the same strategy sets, but different payoffs. We assign the guilt cost some properties that psychological and experimental research has discovered. We are interested in which action profiles are *agreeable*. Agreeability, in turn, is defined in terms of incentive compatibility and individual rationality. An action profile is *incentive compatible* if neither player prefers breaching - for any breach the guilt cost is larger than or equal to the underlying game benefit. An action profile is *individually rational* if each player gets more than in her least preferred Nash equilibrium when neither breaches. More technically, we are interested in the properties of the transformed game when varying (1) the orginal game, (2) the agreement, and (3) how prone to guilt the players are. We remain agnostic to the fine details of the negotiation protocol, and also we suppose that the proneness to guilt types are common knowledge.

We assign the following properties to the guilt cost:

{A} guilt is weakly increasing in the harm that the player causes to the opponent by breaching an agreement

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

{C} guilt is weakly increasing in player's agreed payoff

{D} if no agreement is reached, then there is no guilt cost

These features are intuitively and introspectively appealing. 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. Property {B} is a sort of 'no sucker' property. The player will not feel guilt about breaching an agreement if the opponent breaches the agreement as well. According to property {C} an opponent's generosity and kindness is associated with higher guilt. Since there is guilt only if the opponent does not breach the agreement, the fact that the opponent respects together with the high agreed payoff indicate that the opponent is being kind and genereous towards the player. Not reciprocating this by breaching the agreement will cause more guilt than if the agreement had been less generous. Property {D} expresses the idea that if nothing is promised then there cannot be guilt for not doing as promised. In addition to their intuitive appeal, we present experimental evidence and psychological theory that supports these assumptions in section 2.

As an example, let Ray choose rows and Cal choose columns in a prisoner's dilemma. When they do not communicate, the payoffs are as follows

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{array}{cc} 2,3 & -1,\ 5 \\ 3,-3 & 0,0 \end{array} \end{array} \tag{1}$$

The payoffs are normalised so that, if both defect, zero payoff results to both. For Ray, the payoff difference between both cooperating and both defecting is 2. This difference is Ray's *agreed payoff*. Having agreed on cooperation, if Ray defects but Cal cooperates, Ray's payoff is increased by 1 and Cal's payoff is decreased by 6. We call the former difference Ray's *benefit* from breaching and the latter difference the *harm* that Ray causes to Cal. Our theory implies that Ray's incentive to breach the cooperative agreement is the lower, the higher is Ray's agreed payoff, the higher the harm that Ray inflicts on Cal, and the lower Ray's benefit from breaching.

As for more general results, *each player can agree on any of her individually rational action profiles where she is required to choose an underlying game best reply.* This is obvious, since the guilt cost can only strengthen underlying game incentives. Second, since there is no guilt when the agreed payoff is equal to or below the worst Nash equilibrium payoff, *incentive compatibility implies individual rationality* when off the best reply correspondence.

Furthermore, in games where payoff is concave in each of the two actions, *checking for marginal incentive to breach is necessary and sufficient for incentive compatibility* as long as guilt cost is convex in harm: the harm caused to the opponent is convex as a rescaled negative of opponent's payoff. Thus, checking for the marginal incentive to breach for both is broadly necessary and sufficient also for agreeability.

Next, we consider marginal changes in one or the other of the agreed actions. The marginal incentive to breach has two components: (i) the marginal benefit from breaching and (ii) the marginal guilt cost. The latter is affected by two

factors: (a) the agreed payoff and (b) the harm on the opponent. Marginal changes in the terms of the agreement will have an effect on each of these three: (i) the marginal benefit, (iia) the agreed payoff and (iib) the marginal harm. In supermodular games, where the payoff is increasing in opponent's action, varying an agreed action marginally will have unambiguous effects on player's incentives to breach: First, *increasing opponent's action decreases her marginal incentive to breach.* Second, *increasing player's action increases her marginal incentive to breach*[1]. Furthermore, given that the payoff is concave in each action, it may be convex and increasing in equal changes of both actions in supermodular games only. In such a case, we have the result that *if any underlying game non-equilibrium action profile is agreeable, then an efficient action profile is.* Notice that applications include public good provision games, moral hazard in teams, and some bertrand dupolies with imperfect substitutes. Yet, the monotonicity and efficiency results do not hold for homogenous good cournot duopoly, for instance!

The paper is organized as follows. Section 2 presents related literature in economics and in psychology. Section 3 presents the model. Section 4 discusses the prisoner's dilemma and section 5 studies a public good game. Section 6 presents general results. Section 7 studies a cournot duopoly. Section 8 concludes and discusses some further research problems.

2 Related literature

Economics. Nash [35] already interprets two-player cooperative games as bargaining about strategies to be chosen in a follow-up game where players can enforce any strategies they agree upon - all action profiles are agreeable.

Formal models of communication, information transmission and preplay negotiation, starting from Crawford and Sobel [11] cheap-talk model assume that players can communicate but that messages are not arguments in the utility function. In cheap talk, there is no common knowledge whether an agreement is reached[2]. On the other hand, the agreements are just talk and by no means enforceable. Cheap talk models predict that an action profile must be at least rationalizable or even a Nash equilibrium to be agreeable[3]. Thus, in a public good game where contributing nothing is a strictly dominant strategy, pre-play communication should not affect play.

Yet, evidence from experiments on communication in such games shows that communication does increase effciency and that non-equilibrium outcomes are proposed and agreed upon (Ledyard, [32]). Earliest experiments to show this in the prisoner's dilemma case were Loomis [33] and Radlow and Weidner [39]. Recent studies for the two-person prisoner's dilemma case is provided by Duffy and Feltowich [16], [17]. Extensions to public good provision games have been

[1] This, in fact, holds for any game with concave and increasing payoffs in opponent's action.

[2] In the purest model, the fact that language is not common knowledge implies this if rationalizability is the appropriate solution concept.

[3] Aumann [3] argues that the set of possible agreements is even smaller.

considered and the robustness of this result is verified by various experiments, for instance, Dawes, McTavish, and Shaklee [15], Isaac, McCue, and Plott [30], and Isaac and Walker [31].

Despite this overwhelming evidence, there have been few attempts to explain the reasons lying behind the phenomenon. Exceptions include both the reciprocity theory , (Rabin [38]) and inequity aversion theories (Fehr and Schmidt's [22], Bolton and Ockenfels [9]). These theories can account for the experimental findings[4] as long as the payoffs are not too unequal.

Nevertheless, Charness and Dufwenberg [10] (CD) on the one hand[5], and Gneezy [25] on the other hand, carry out further communication experiments and show that neither of the above mentioned theories can fully account for the detected behavioral patterns. They conclude that there must be a separate preference related to lying. CD present a stylized model of let down aversion in the context of a sequential prisoner's dilemma where a player suffers a cost when she acts counter to the opponent's expectations on her behavior[6]. In their model, promising to carry out an action is assumed to strengthen the belief that the opponent expects corresponding behavior thereby creating further incentives to behave accordingly. Also implicit in their model is an assumption that guilt is an increasing function of the expected harm caused to the opponent. Nevertheless, the role of communication is only implicit in their model. There is no reason why beliefs should be affected in the prescribed manner, and more importanty, if anything else affected the beliefs in an equivalent manner as communication, the results should be the same.

In the present model, players do not dislike acting counter to expectations of the opponent but counter to an agreement. We are explicit not only about the effect of communication and the agreement, but also about the effect of harm that affects guilt. The model is general. It captures many features of reciprocity, yet avoiding problems of tractability in models where payoffs depend on beliefs explicitly[7]. The guilt in our model bases its properties on research in social psychology and allows for most of the features relevant to pre-play negotiation. The model should be regarded as complementary to other social preference models.

Guilt has been discussed in several papers since Frank [23] who argues that it may well be materially profitable for an agent to have a conscience, dislike for disobeing social norms. A recent model on emotional cost of breaching social norms is provided by Huck, Kubler, and Weibull [29]. These models involve no communication. Ellingsen and Johanneson [19] do allow for communication and study the interplay of inequity aversion and guilt in a hold-up problem setup

[4] For the former, the theory of sequential reciprocity must be applied. The latter, must be combined with the ideas of Farrel (1987) which supposes that a pre-play message is followed as long as there is no incentive not to do so.

[5] See also Dufwenberg [18]

[6] Thus, like the theories of reciprocity, the theory falls into the category of psychological game theory (Geanokoplos, Pierce, and Stachetti [24]) where players' payoffs depend on beliefs explicitely.

[7] Some feasible guilt cost functions imply that the preferences in the cases where an agreement is in place are tractable social preferences of Cox and Friedman [14].

between a seller and a buyer. A seller invests in a good with certain and positive net returns but the buyer has the bargaining power to propose a division of gross returns which the seller can only accept or reject. Rejection leads to the loss of return to the investment. So as to the communication, a seller can make threats of rejecting some unfair proposals and the buyer can make promises to make sufficiently fair proposals. Their model is similar to ours in that guilt does not depend on the beliefs explicitely. As well, guilt is suffered if one breaches a promise. However, their model of guilt is simpler, since it assumes that breaching a non-binding agreement imposes a constant guilt cost. They find that fair-mindedness strengthens the credibility of promises to behave fairly by the buyer, but weakens the credibility of threats to punish unfair behavior by the seller.

Psychology. In addition to their intuitive appeal, properties {A} to {D} are supported by experimental evidence and psychological theory. So as to property {A}, Hoffman [27] suggests that guilt has its roots in a distress response to the suffering of others. Okel and Mosher [36] find that subjects feel more guilty about derogation, the more pronounced the impact seems to be on the victim. The main empirical finding of Gneezy [25] is that "lying is directly costly" and that "people do not care only about their own gain from lying; they are also sensitive to the harm that lying may cause to the other side.

As far as property {B} is concerned, Baumeister, Stillwell, and Heatherton [5] find that people felt more guilty about transgressions involving an esteemed person than about transgressions involving someone they held in low regard. It is rather appealing to suppose that, if the opponent breaches the agreement, the esteem of a guilt-prone player towards the opponent is smaller than if the opponent respects. We go to an extreme and assume that the player does not suffer from guilt if the opponent breaches the agreement.

Property {C} operates together with property {B}: agreements that are respected and give a high payoff to a player, signal opponent's concern for player's welfare and such opponent's are likely to be esteemed. According to Clark and Mills [12] and Clark [13], concern for the other's welfare is the defining feature of communal relationships as opposed to exchange relationships. According to this research, guilt is more likely to arise in the former type than in the latter type of relationships. First, if the opponent respects the agreement, she shows more concern for the player than if she does not. But not only is respecting an agreement an indication of concern for the opponent. In addition, such a concern is signalled by allowing an agreement where the player gains more conditional on respecting. An opponent that lets the player to have a high payoff by respecting and by assigning a high payoff to her in the agreement, is esteemed by the player, and thus the player's guilt is higher if she breaches but the opponent does not.

So as to property {D}, negotiation that ends up in an agreement explicitely states an expectation and a standard of behavior for the play phase. Not reaching an agreement indicates inconcurrence of such a standard among the players. Millar and Tesser [34] note that guilt depends on a 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. On the other hand, many experimental studies of the public good game show that a single message for not contributing is sufficient to make an agreement invalid.[8] This body of reserach suggests each player should have an ability 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.[9]

More generally, research in psychology identifies three types of emotional distress associated with lying: guilt, shame, and fear of punishment. From a game theoretical perspective, the latter two have a reputational and repetitional flavour respectively whereas guilt may be suffered even if the act of lying is unobservable and unverifiable to others, or the victim or a third party is in no position to retaliate.

According to Baumeister, Stillwell, and Heatherton [4], "guilt can be distinguished from fear of punishment on the basis that the distress pertains to the action itself rather than to the expectation of hedonically aversive consequences of the action. ...One can clearly feel guilt..., even if the victim is in no position to retaliate."

Baumeister, Stillwell, and Heatherton [4] are concerned with "what makes people feel guilt and what that feeling - or the motivation to avoid that feeling - causes them to do" (p.245). They argue that:

> "From an interpersonal perspective, the prototypical cause of guilt would be the infliction of harm, loss, or distress on a relationship partner. Although guilt may begin with close relationships, it is not confined to them; guilt proneness may become generalised to other relationships. ... In particular, a well-socialized individual would presumably have learned to feel guilty over inflicting harm to even a stranger."

In the present model, as in theories of fairness, players internalise the opponent's payoff *but only conditional on reaching an agreement*, *conditional on the opponent respecting the agreement* and *conditional on the opponent suffering from breaching*. Thus, the model shares some of the features of the models of fairness but differs from those in important dimensions.[10]

3 The Model

Let Γ be a two-player simultaneous move normal form game, below referred to as the *underlying game*[11]. Before the game is played, the two players negotiate

[8] See Ledyard [32] and Shankar and Pavitt [37].

[9] From the perspective of the functional classification of communication of Bornstein [8], our model is able to capture trust-enhancing {C}, norm-emergence {D}, and agreement-formation {D} functions of communication. These explanations for why communication matters are supported and prioritized over other explanations by a bulk of experimental evidence. See Shankar and Pavitt [37] for a survey.

[10] Gneezy [25] provides direct evidence that guilt cannot be captured by the fairness models alone.

[11] We rule out the use of mixed strategies in the underlying game. Sticking to the literary interpretation of a mixed strategy, supposing that the player feels guilty when using a different probability distribution over actions than the one agreed would violate the structural premise that payoffs are defined on the set of consequences. On the other hand, the breaching cost in

abouth which actions to choose. Generally, the negotiation may have an arbitrary strategic structure with the only requirement that it ends in an agreement on how to play, $m = (m_1, m_2)$, or disagreement[12], m_o.

3.1 The underlying game

The two-player *underlying game* is given by $\Gamma = \{S_i,\ u_i(s) : S \to R\}$. The action set of player i is S_i. A combination of actions is an *outcome* $s = (s_i, s_j) \in S = S_i \times S_j$. The *underlying game payoff* of player i is $u_i(s)$.

The lowest Nash payoff of player i is $u_i^* \equiv \min_{s \in NE(\Gamma)}\{u_i(s)\}$ where $NE(\Gamma)$ is the set of Nash equilibria in the underlying game with $u^* = (u_i^*, u_j^*)$. If rational players play without preplay negotiation and they have correct expectations about the behavior of the other, then a Nash equilibrium should result. Thus, the lowest Nash payoff is the worst case scenario for failure of negotiation for each player.

We restrict $m \in S \cup \{m_o\}$. Thus, negotiation ends up in an agreement on how to play or in disagreement. 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* is the payoff that the player gets if both respect the agreement, $u_i(m)$. If player i deviates from the agreement, the payoffs will generally be affected and the suffered *harm* is j's payoff difference between the agreed action profile and the outcome that would result from i's deviation, $h_j(m, s_i) \equiv u_j(m) - u_j(m_j, s_i)$. Similarly, i's *benefit from breaching* is $b_i(m, s_i) \equiv u_i(m_j, s_i) - u_i(m)$.

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*, $g_i(u_i(m), h_j(m, s_i))$, depends on how much harm she inlifcts on her opponent and how nicely i herself is treated in the agreement.

The utility function over the outcomes in the entire game is assumed to be additively separable in guilt and the underlying game payoff.

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

The entire game payoff now depends on m and, due to guilt, talk is not cheap. The guilt cost is represented by $\theta_i g(u_i(m), h_j(m, s_i))$ which is assumed to be non-negative. This rules out revengeful feelings or spite, on the one hand,

our model can be interpreted both as shame or as guilt as long as the actions are observable. If a player chooses a mixed strategy, it may be argued that the player has no reason to feel shame if an action in the support of the agreed mixed strategy is drawn even if the distribution in use differs from the agreed probability distribution. There may be guilt even if there is no shame.

[12] *Preplay negotiation* is a finite extensive form game tree. The terminal histories are associated with an oral (non-binding) agreement, $m = (m_i, m_{-i})$ or with disagreement, m_o.

and positive emotions related to respecting agreements, on the other hand[13]. The parameters $\theta = (\theta_1, \theta_2)$ captures players' *pronenesses to guilt*. For a given deviation, the players with a higher proneness to guilt suffer more. We only alloaw 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 is one of cheap talk.

Notice first, that the guilt cost depends on the agreement and 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 (guilt cost is zero).

Furthermore, (BD) implies that if disagreement is reached, then there is no guilt cost. We assume that each player can unilaterally enforce disagreement[14], m_o. Also, there are no bad feelings about own cheating if the opponent cheats too. Thus (BD) incorporates properties {B} and {D} into the guilt cost.

Furthermore, we assume, that the guilt cost $g(u_i, h_j)$ is weakly increasing in u_i and in h_j. The more harm a player causes by not doing as agreed and the more nicely the opponent treats the player in the agreement, the more the player suffers from guilt. Condition AC introduces properties {A} and {C} into the guilt cost.

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

Obviously, if the guilt function is differentiable then these monotonicity properties would simply amount to positive derivatives, $\frac{\partial g}{\partial u_i} \geq 0$ and $\frac{\partial g}{\partial h_j} \geq 0$.

Also, we assume that if the player causes no harm to the opponent[15] 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 caused and the agreed payoff is strictly above the worst Nash payoff, then the guilt cost is strictly positive:

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

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

$$g(u_i, h_j) = \left\{ \begin{array}{c} \gamma \text{ if } h_j > 0,\ u_i > u_i^* \\ 0 \text{ otherwise} \end{array} \right. , \tag{2}$$

[13] This is somewhat restrictive. In a more general setup, one could assume that the player may feel good about respecting the agreement or feel good about breaching the agreement if one is treated badly in communication or in play.

[14] This property rules out any emotional factors driving behaviour when players fail to reach an agreement. This is restrictive. For instance, disappointment or anger related to the fact that an opponent has enforced a disagreement might well affect behavior.

[15] Andreoni (2005) provides some indirect evidence for this. In his extension of the buyer-seller trust game where sellers can make non-binding promises of refunds, the sellers who promise a refund, increase the return rates (quality) above no-buy utility so that no harm is caused, if a promised refund request is rejected. Thus, for any realised rejection of refund, guilt is avoided, and the present theory predicts rejection conditional on refund request and return rate above one which the data in Andreoni seems to confirm.

or guilt cost that only depends on one of the arguments is allowed.

Another example of a guilt cost function with all the properties assumed in this section is [16]

$$g(u_i(m), h_j(m, s_i)) = \max\{h_j(m, s_i), 0\} \max\{u_i(m) - u_i^*, 0\} \tag{3}$$

This function is zero if the harm on the opponent is non-positive or if the agreed payoff is below u_i^*. Otherwise, it is strictly positive. It is increasing in the harm inflicted and in the agreed payoff.

Let us now introduce some further notation. Denote by $BR_i(s_j)$ the underlying game best reply correspondence of player i. Denote by $\Gamma(m;\theta)$ a subgame where m is agreed when pronenesses to guilt are θ and by $s^*(m;\theta) = (s_i^*(m;\theta), s_j^*(m;\theta))$ the equilibrium correspondences in that subgame. Given an arbitrary game Γ', we denote by $NE(\Gamma')$ its set of Nash equilibria.

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

$$U_i(m_i, m_j, s_i, m_j) = u_i(m) + b_i(m, s_i) - \theta_i g(u_i(m), h_j(m, s_i)) \tag{4}$$

and

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

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

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 is guaranteed for both players, the agreement m is a Nash equilibrium of the subgame where m is agreed upon, $\Gamma(m;\theta)$. On the other hand, an agreement m is called *individually rational* if neither prefers to enforce disagreement rather than agreeing on m when both respect the agreement

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

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

We now define *the agreeable set of player* i as $A_i(\Gamma, \theta_i) \equiv \{m|\ m \text{ satisfies } (IC_i) \text{ and } (IR_i)\}$ and *the agreeable outcome set* is defined naturally as the intersection of the two agreable sets, $A(\Gamma, \theta) \equiv \cap_{i=1,2} A_i(\Gamma, \theta_i)$. We call an action profile in the agreeable outcome set simply *agreeable*.

[16] The entire game preferences of this form belong to the class of Fox-Friedman [14] preferences with $\alpha = 1$ with the emotional state depending on the assigned payoff $u_i(m) - u_i^*$.

4 Prisoner's Dilemma

Let us reconsider the prisoner's dilemma in (1). Let us write the game in more general terms as

$$\begin{array}{cc} & \begin{array}{cc} C & \qquad\qquad D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{array}{cc} u_1, u_2 & \qquad u_1 - h_1, u_2 + b_2 \\ u_1 + b_1, u_2 - h_2 & \qquad 0, 0 \end{array} \end{array} \tag{6}$$

We suppose that the guilt cost takes the simple form of the example given in (3).

Given the parameters of the model, we are interested whether agreeing on cooperation can be sustained as a subgame perfect equilibrium. Given such an agreement, player i's payoff from playing according to the agreement is u_i. Given that the opponent respects, player i respects the agreement if

$$u_i \geq u_i + b_i - \theta_i u_i h_j$$

Or equivalently,

$$\theta_i \geq \frac{b_i}{u_i h_j} \tag{7}$$

This is an incentive compatibility condition for both cooperating. Moreover, cooperation is individually rational by the structure of the prisoner's dilemma game. So, an agreement on (C, C) should be particularly easy to reach if b_i is small and h_j is large, as Gneezy [25] suggests. Also, a large u_i facilitates cooperative agreements. This gives us comparative statics results that are testable.

In the prisoner's dilemma, individual rationality rules out patterns (C, D) and (D, C). Both defecting is incentive compatible and individually rational for all types since it is the unique nash equilibrium. Hence, (D, D) is always agreeable and (C, C) is agreeable if the incentive to breach is non-positive for both.

Proneness to guilt may transform a prisoner's dilemma into a coordination game. This is a familiar property from fairness models. Yet, here the ability to commit to cooperate does not depend on how much more or less the opponent gets when players cooperate, $u_i - u_j$. It depends on how much more the player gets when players cooperate than when players defect, $u_i - 0$. On the other hand, the payoff of the opponent is internalised only to the extent of how much player's defection affects the opponent's payoff.

Guilt imposes a cost of defection. It is almost trivial that if this cost if sufficiently large to overweight the benefit from breaching, the player can credibly commit not to defect. Yet, the prisoner's dilemma is a rather degenerate game in the class of games with inefficient equilibria where preplay negotiation insight is expected to be particularly valuable. There is only one action profile that Pareto dominates the underlying game Nash equilibrium and the set of agreements under negotiation is very limited.

5 A Public Good Game

A game to which we can easily generalise the prisoner's dilemma type of argumentation is the following linear public good game. Each player has an endowment of ten dollar coins. Each player decides how many dollars to contribute, $s_i \in \{0, ..., 10\}$. The payoffs are given by.

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

We suppose that $\frac{1}{2} < \alpha < 1$ so that the game has the linear public good structure. Here $\alpha < 1$ is the *marginal per capita return* (MPCR). Hence, in the unique Nash equilibrium, both contribute zero. The *marginal group return* equals 2α. Therefore, it is socially optimal that both contribute everything they have.

We suppose that the guilt cost is given by (3). Players can agree to any agreement where both get a positive payoff and the guilt is sufficient to prevent breaching. The benefit from a unit underprovision vis à vis the agreement is $1 - \alpha$ and the harm is α. Notice first, that due to the linearity of payoffs, it is sufficient to check for one unit underprovision only. This property applies for a larger class of games as we will show in section 6. Let us call $1 - \alpha - \theta_i \alpha u_i(m)$ player's *marginal incentive to breach*.

A player will contribute zero dollars if and only if a unit underprovision is beneficial. Hence player i can agree on any agreement where

$$\theta_i \geq \frac{1 - \alpha}{\alpha \max\{(\alpha(\sum_{k=1,2} m_k) + 10 - m_i), 0\}} \tag{8}$$

and where

$$\alpha(\sum_{k=1,2} m_k) + 10 - m_i \geq 0. \tag{9}$$

The first condition (8) is an incentive compatibility condition whereas the second (9) is an individual rationality condition. On the one hand, the cost of breaching must dominate its benefit. On the other hand, each player must get at least the Nash equilibrium payoff in order to find negotiation worthwhile.

Notice that individual rationality condition is actually redundant: it is implied by the incentive compatibility condition. When agreed payoff approaches zero, the RHS of (8) approaches infinity and the agreement is not incentive compatible for any type.

Some of the properties explicit in (8) are worth emphasizing. First, the relative activity, $m_1 - m_2$, matters but not the relative payoffs, $u_1(m) - u_2(m)$. Only player's own payoff, $u_i(m)$, enters the inequality, not the opponent's payoff, $u_j(m)$. The incentive to breach is deacreasing in the opponent's contribution, m_j, but increasing in the player's own contribution, m_i, in (8). For a given contribution of the opponent, the more player i is required to contribute and the less player j is required to contribute, the less likely it is that i can agree

on m. Furthermore, a player with a higher proneness to guilt can agree on a larger set of agreements.

Specific to the game is the effect of the marginal per capita return, α. This decreases the benefit from breaching $1-\alpha$ and increases the harm inflicted on the other α. Furthermore, for a larger α, the agreed payoff of any agreeable action profile is higher vis à vis the underlying game equilibrium where the payoff is 0. All these effects make it easier for the players to agree when the marginal per capita return is higher. Hence, increasing α makes the incentive compatibility constraint less stringent for each agreement. But, perhaps most importantly, socially better agreements are easier to enforce, since the incentive compatibility constraint becomes less stringent in the sum of contributions $\Sigma_{k=1,2}m_k$.

Let us collect the findings in this section into a proposition.

Proposition 1 *Let g be convex in h_j. In the public good game,*

- *an agrement is incentive compatible iff marginal incentive to breach is non-positive for $i = 1, 2$*
- *player i's marginal incentive to breach is decreasing in α and in m_j and in $\sum_{k=1,2} m_k$ and increasing in m_i*

Proof. It is straightforward that m satisfies IC_i for $i = 1, 2 \Leftrightarrow m$ is agreeable since IC_i implies IR_i. Thus it suffices to show that non-positive marginal incentive to breach is equivalent to non-positive incentive to breach to all $s_i \in S_i$. We have for all $s_i < m_i$

$$
\begin{aligned}
& (1-\alpha)[m_i - s_i] - \theta_i \frac{\partial g}{\partial h_j}(u_i(m), h_j(m, s_i))\alpha[m_i - s_i] && (10)\\
\leq\ & (1-\alpha)[m_i - s_i] - \theta_i \frac{\partial g}{\partial h_j}(u_i(m), h_j(m, m_i))\alpha[m_i - s_i] && (11)\\
\leq\ & 1 - \alpha - \theta_i \frac{\partial g}{\partial h_j}(u_i(m), h_j(m, m_i))\alpha && (12)\\
\leq\ & 0 && (13)
\end{aligned}
$$

where the first inequality follows from the fact that opponent's payoff is increasing in s_i and that g is convex in h_j, and the second inequality follows from the fact that $[m_i - s_i] \geq 1$. ■

Figure 1 describes the agreeable outcome set for $\alpha = \frac{3}{4},\ \theta_i = 4$.

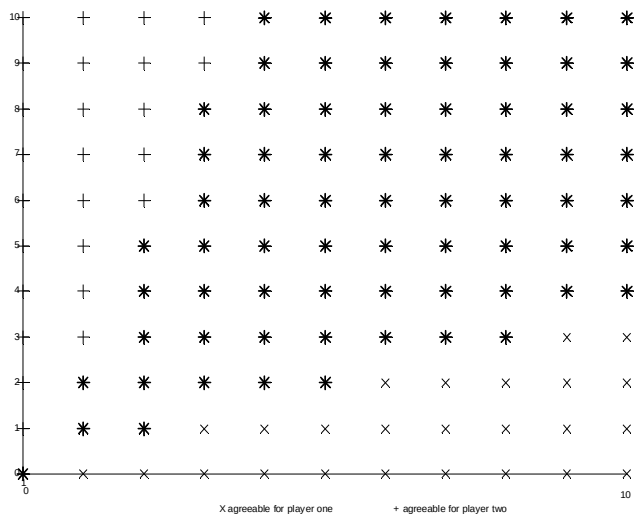

Figure 1: The Agreeable Set

The underlying game best reply curves lie on the axes. The outcomes that are agreeable for player one are marked with plus signs and the outcomes that are agreeable for player two are marked with crosses. Thus the outcomes marked with asterisks are agreeable outcomes, $A(\Gamma_{PG}(\frac{3}{4}),(4,4))$. Notice, that the best reply curves lie on the axes and that each player's best reply curve is agreeable for each player. Thus, the Nash equilibrium play, $(0,0)$, is agreeable. Notice also that some efficient outcomes are agreeable, for instance the symmetric efficient outcome where both give a full contribution, $m=(10,10)$. Figure 2 studies the agreeability of the symmetric efficient outcome. There, the treshold proneness to guilt that is indifferent between breaching and respecting the symmteric efficient agreement is mapped as a function of α. As stated above, increasing α makes the incentive compatibility constraint less stringent and, thus, the function is decreasing.

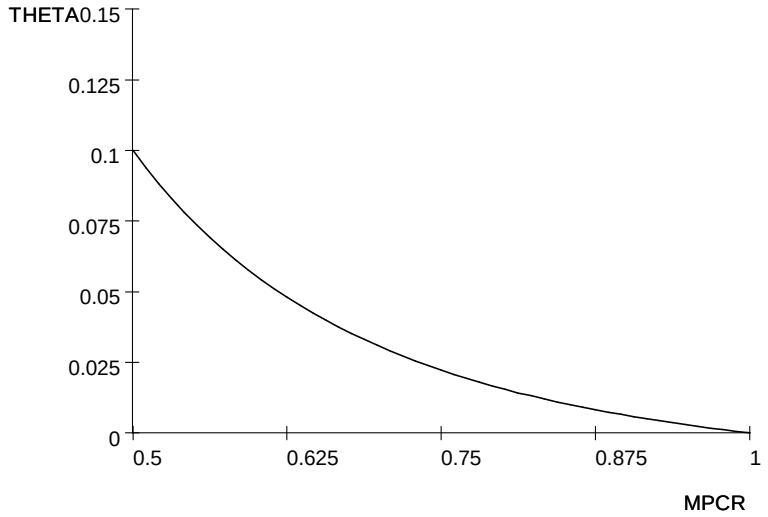

Figure 2: Indifferent player at efficiency as a function of α

6 General Properties of the agreeable outcome set

This section derives some simple properties that apply to any normal form underlying game. The property that each player can agree on every outcome where the player's agreed action is the underlying game best reply to the agreed action of the opponent is a very robust property. Underlying game preferences drive the player to choose the best reply. The guilt cost can only strengthen this incentive, since other actions may be associated with guilt.

Lemma 1 *Let* $m_i \in BR_i(m_j)$. *Then* $m \in A_i(\Gamma, \theta_i)$ *iff* $u_i(m) \geq u_i^*$

Proof. See appendix. ■

Notice, that in the prisoner's dilemma example, the defection equilibrium is agreeable for any proneness to guilt types. For zero proneness to guilt types, this is the only agreaable action profile whatever the payoffs are in the prisoner's dilemma game. In general in cheap talk, the agreeable outcome set is just the set of Nash equilibria of the underlying game. A Nash equilibrium is a action profile where each player is best replying to the action of the opponent. Thus, a Nash equilibrium belongs to the agreeable outcome set of each player. Hence, it belongs to the agreeable outcome set.

Proposition 2 *If* $m \in NE(\Gamma)$, *then* $m \in A(\Gamma, \theta)$.

Proof. See appendix. ■

Guilt never reduces the menu of agreements available to the players. To the contrary, the public good example shows that positive proneness to guilt can dramatically increase the set of plays that are agreeable.

Yet, preplay negotiation 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. If both players defect, neither feels guilt and payoffs involve only underlying game payoffs. This insight is easily generalized: it is straightforward that an underlying game equilibrium where neither respects the agreement m is an equilibrium of the subgame $\Gamma(m;\theta)$ for all proneness to guilt types. This shows that even if m is a Nash equilibrium of $\Gamma(m;\theta)$, there may be other equilibria as well.

Lemma 2 *If for* $i = 1, 2$, $m_i \not= s_i^*$ *and* $s^* \in NE(\Gamma)$ *then* $s^* \in NE(\Gamma(m;\theta))$

Proof. See appendix. ■

Lemma (2) establishes that players' proneness to guilt may create or aggravate the coordination problem involved in the multiplicity of equilibria. Farrell [20] refines the Nash equilibrium concept in the subgame $\Gamma(m;\theta)$ by assuming that if m is a Nash equilibrium of $\Gamma(m;\theta)$, then m will be played, $s^*(m;\theta) = m$. This refinement makes truth focal: players will conform to the agreement, if there is no incentive not to do so.[17] If we apply Farrell's refinement, then players will always play according to an agreement if it is agreeable.

Notice yet, that it is not true that an underlying game Nash equilibrium is an equilibrium after any agreement. Nash equilibria may be removed from the game. Consider the following game of chicken

$$\begin{array}{ccc} & L & R \\ T & -1,-1 & 2,0 \\ B & 0,2 & 1,1 \end{array} \tag{14}$$

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 (2) with $\gamma = 1$. Let us suppose that players agree

[17] There is an idea of a common language implicit in this assumption. The meaning of a message or an agreement is common knowledge and this allows each player to verify whether some player has an incentive to deviate from the agreement if others respect the agreement.

Self-commiting agreements are such that, for both players, if the player knows that the opponent believes that the player plays according to the agreement, then it is optimal for the player to choose her agreed action

Farrell and Rabin [21] discuss messages that are self-enforcing. There are three reasons to be suspicious about a message (or an agreement). First, players may have different understanding what the message means. Second, even if messages are understood correctly, players may have incentives to mislead their opponents. Self-signalling messages are sent, if and only if they are true. Self-commiting messages are such that if believed, the sender will have an incentive to do accordingly. Aumann [3] presents a simple game where the self-signalling condition is violated. We shall consider agreements that satisfy the self-commiting condition for each player as agreeable.

on playing (B, R) which gives agreed payoff 1 for player one. Now, if player one breaches the agreement and chooses T instead, she gets $2 - 2 = 0$ which is smaller than 1 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, we show that an agreement where one of the players can make both better off by deviating unilaterally from the agreement (even if the opponent respects the agreement) does not belong to the agreeable outcome set.

Lemma 3 *For any m, if there is a player i such that there is s_i such that $u_i(s_i, m_j) > u_i(m)$ and $u_j(m_j, s_i) \geq u_i(m)$ then $m \notin A(\Gamma, \theta)$ for any θ.*

Proof. See appendix. ■

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

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

$$\begin{array}{cc} L & R \\ T \quad 2,2 & 0,100 \\ B \quad 1,1 & 1,1 \end{array} \tag{15}$$

since if player one breaches and chooses T, both players are better off. One could argue that 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. But of course, player one would then be inclined to choose B. 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 [20]. Indeed, this type of plurality may question whether (B, L) is agreeable in the first place. For our analysis, it is sufficient to notice that since player 1 can make both better off, the agreement is not incentive compatible.

In (15), players cannot agree on (T, R) either since player one 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 cannot agree on (T, L) either due to player two's high gain from choosing R instead. But if we let player two's proneness to guilt to become sufficiently high, (T, L) becomes agreeable. As the proneness to guilt becomes infinite, the guilt cost becomes infinite for deviations that cause a positive harm. Hence, whenever deviation causes harm, it will not be made. In general, as the players' proneness to guilt becomes infinite, the players can agree on any individually rational play for which there is no unilateral pareto-improving deviation.

Proposition 3 *Suppose that if $u_i(m) > u_i^*$, g is strictly increasing in h_j. Let the underlying game payoffs be finite. Let $u_i(m) > u_i^*$ 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 $u_j(m_j, s_i) < u_i(m)$*

Proof. See appendix. ■

Lemma 3 has another implication. Namely, within the agreeable outcome set, the interests of the players are opposed for any change of one of the agreed actions only

Proposition 4 *Let* $(m_i, m_j), (m_i', m_j) \in A(\Gamma, \theta)$ *then*

$$\begin{aligned} u_i(m_i, m_j) > u_i(m_i', m_j) &\Rightarrow u_j(m_i', m_j) > u_j(m_i, m_j) \\ u_j(m_i', m_j) > u_j(m_i, m_j) &\Rightarrow u_i(m_i, m_j) \geq u_i(m_i', m_j) \end{aligned} \tag{16}$$

Proof. See appendix. ■

6.1 Finite Games with Ordered Strategy Spaces

Let us now focus on finite games with *ordered* strategy spaces, $S_i = \{s_{i,1}, ..., s_{i,n}\}$. Inspired by the results in the public good game where actions are ordered in terms of contributed amounts, we seek to generalise two results gained there: first, that the marginal incentive to breach condition is necessary and sufficient for an outcome to be agreeable; second, that the marginal incentive to breach is monotone in each of the agreed actions. We show that, when the guilt cost is convex in the harm, the first results generalises to underlying games with concave payoff functions in each action. The second result holds in these games but, in addition, the actions must be strategic complements.

Furthermore, we establish an efficiency result that is also true in the public good game: an efficient agreement can be made if any non-underlying game agreements can be made. This is true for games where the actions are not strategic subsitutes and where certain convexity properties hold.

We first set the scene by making further assumptions on the underlying game payoff and the guilt cost.

ASSUMPTIONS ON THE UNDERLYING GAME PAYOFF, $u_i(s)$

In addition to supposing that the game is finite, we suppose that

{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:

$$u_i(s_{i,k}, s_j) - u_i(s_{i,k-1}, s_j) - [u_i(s_{i,k-1}, s_j) - u_i(s_{i,k-2}, s_j)] = \delta \leq 0$$

and

$$u_j(s_j, s_{i,k}) - u_j(s_j, s_{i,k-1}) - [u_j(s_j, s_{i,k-1}) - u_j(s_j, s_{i,k-2})] = \sigma \leq 0$$

{3} the payoff functions are supermodular

$$u_j(s_{j,l}, s_{i,k}) + u_j(s_{j,l-1}, s_{i,k-1}) - [u_j(s_{j,l}, s_{i,k-1}) + u_j(s_{j,l-1}, s_{i,k})] = \phi \geq 0 \tag{17}$$

These properties are satisfied in the public good game, but in a degenerate manner: $\delta = \sigma = \phi = 0$.

ASSUMPTIONS ON THE GUILT COST, $g(u_i, h_j)$

In addition to assumptions on the underlying game, we assume that the guilt cost function is convex in the harm, h_j, and in the agreed payoff, u_i, and that it is supermodular in its two arguments

{4} g is convex in h_j

{5} g is supermodular and convex in u_i

Notice that the fact that the payoff is concave in 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. Thus, by assumption {4}, the guilt cost is convex in s_i as a combination of two convex functions. On the other hand, the underlying game payoff u_i is concave in s_i. Consequently, the problem of choosing the optimal deviation given that the opponent respects is a simple convex optimisation problem. Hence, checking that neither preferes to breach the agreement marginally is necessary for an agreement to be incentive compatible.

To simply formulate such a condition, we extend the concept of marginal incentive to breach from the public good game example:

Definition 1 *(Marginal incentive to breach)*

If $u_i(m_i - 1, m_j) - u_i(m) \geq 0$

$$\beta_i(m, \theta_i) \doteq b_i(m, m_i - 1) - \theta_i g(u_i(m), h_j(m, m_i - 1))$$

If $u_i(m_i + 1, m_j) - u_i(m) > 0$

$$\beta_i(m, \theta_i) \doteq b_i(m, m_i + 1)$$

The fact that $\beta_i(m, \theta_i)$ does not involve any guilt cost when $u_i(m_i+1, m_j) - u_i(m) > 0$ is due to assumption {1}: player one does not make player two worse off and thus does not suffer from guilt. Consequently, assumption {1} on the underlying game payoffs together with lemma 3 gives us a necessary condition for an action profile to be agreeable. The play must belong to the following set[18]

$$M_F = \{m | u_i(m_i, m_j) \text{ is non-increasing in } m_i \text{ for } i = 1, 2\} \tag{18}$$

Let *the boundary of* M_F be defined as follows

$$\underline{M}_F = \{m | m \in M_F \text{ and there is } i \text{ such that } (m_i - 1, m_j) \notin M_F \text{ or } (m_i + 1, m_j) \notin M_F\} \tag{19}$$

Thus, the boundary of M_F is the set of action profiles such that there is a profile outside M_F reachable by a change of one unit of only one of the actions.

Non-positivity of the marginal incentive to breach is a necessary condition for agreeability. More strongly, that it holds for both, is both necessary and sufficient. First, due to the convexity of the problem if there is no incentive to breach the agreement in the margin, there is no incentive to breach whatsoever.

[18] Except for $m_i = s_{i,n}$ of course.

Second, incentive compatibility implies individual rationality when off the underlying game best reply curves, since if individual rationality is violated the guilt cost is zero and thus incentive to breach is non-positive (zero) only if the agreement is a best reply to the opponent's action.

Proposition 5 *Let Γ be finite. Let $m_i \neq s_{i,1}, s_{i,n}$ and let $m_i \neq BR_i(m_j)$. Let {1}, {2} and {4} hold. Then an action profile is agreeable if and only if the marginal incentive to breach is non-positive.*

Proof. See appendix. ■

In the public good game, we found that the marginal incentive to breach is monotone in each agreed action. We can generalise this property. Let us first consider the effect of m_i on b_i and on h_j. It is necessary that an agreeable outcome lies in M_F. But within M_F, player's payoff must be decreasing in her action. Thus, the effect of player's agreed action on marginal benefit from breaching is nothing but the negative of the second derivative. Thereby, increasing player's agreed action increases her marginal benefit from breaching. Similarly, the effect of m_i on h_j is simply the second derivative, σ, since the harm is itself a rescaled negative of u_j and breaching takes place downwards. Notice then that increasing m_i increases b_i and decreases h_j and thus both these effects have a positive impact on the marginal incentice to breach.

The effects of m_j on b_i and h_j rest on the strategic complementarity of actions: if the opponent increases her action, the player has a stronger incentive to increase her own action. Since breaching takes place downwards, increasing opponent's action dampens the underlying game benefit from breaching. On the other hand, the higher is opponent's action, the more harm decreasing one's action marginally causes to her. Strictly supermodular games, where $\phi > 0$, constitute a set of games where such complementarities are present.

Lemma 4 $h_j(m_i, m_j + 1, m_i - 1) - h_j(m_i, m_j, m_i - 1) = \phi$
$h_j(m_i + 1, m_j, m_i) - h_j(m_i, m_j, m_i - 1) = \sigma$
$b_i(m_i, m_j + 1, m_i - 1) - b_i(m_i, m_j, m_i - 1) = -\phi$
$b_i(m_i + 1, m_j, m_i) - b_i(m_i, m_j, m_i - 1) = -\delta$

On the other hand, proposition 4 together with {1} imply that the agreed payoffs change monotonically in the agreeable outcome set: increasing own agreed action decreases the agreed payoff and increasing opponent's action increases payoff. Thus, also the agreed payoff effect has a positive impact on the marginal incentive to breach when m_i is increased; and a negative impact when m_j is increased.

Thus, in supermodular games, increasing opponent's action decreases the marginal benefit from breaching, increases the marginal harm, and increases the agreed payoff[19]. Thus, unambiguously, the marginal incentive to breach decreases. Similarly, increasing the own agreed action, increases the marginal incentive to breach.

[19] Of course, supermodularity of g is needed so that the interplay between the agreed payoff and the harm effect in the guilt cost does not contradict other effects.

Proposition 6 *Let Γ be finite. Let the actions be ordered. Let {1}, {2}, {3} ,{4} and {5} hold. Then*

$i's$ marginal incentive to breach is increasing in m_i and decreasing in m_j in the agreeable outcome set.

Proof. See appendix ■

The public good example game has another property that can be generalised: a non-equilibrium action profile can be agreed upon if and only if an efficient action profile can be agreed upon.

We suppose that the payoff is convex in symmetric changes of both actions. When the payoff is convex in this way and the payoff is increasing in such changes, it is increasing in symmetric changes up to the point that one of the actions cannot be increased any futher. In such a case, we can use supermodularity and the fact the payoffs are increasing in the opponent's action to argue that action profiles where one of the players chooses her maximal action are efficient. Thus, an efficient action profile is agreeable if any. Consequently, we can argue that an non-equilibrium action profile is agreeable if and only if an efficient action profile is. Yet, we need the non-equilibrium action profile to lie strictly within M_F and not on its boundary.

Theorem 7 *Suppose that $\sigma + 2\phi + \delta \geq 0$. Suppose that the game is symmetric and that {1}, {2}, and {3} hold. Furthermore, let g satisfy {4}, and {5}. Let s^* be the unique equilibrium of the game and $b_i(s_i^*, s_j^*, s^* - 1) = 0$. Then an efficient action profile s^c, such that $s^c \notin \underline{M}_F$ and $s_i^c \geq s^*$ for $i = 1, 2$, is agreeable if and only if some action profile s, such that $s \notin \underline{M}_F$, $s_i > s^*$, $i = 1, 2$, is agreeable.*

Proof. See appendix.

Proposition 6 shows that in games with strategic complementarities the marginal incentive to breach has intuitive monotonicity properties: as the action of the opponent is increased, player's incentive to breach decreases whereas the opposite is true when player's own action is increased.

On the other hand, in supermodular games, broadly speaking, players are able to reach efficient agreements if anything else than underlying game equilibria if the payoff is convex in identical changes in actions.

In addition to the public good game studied above, examples of supermodular games of this type include for instance, moral hazard in teams (Holmström [28]), or bertrand duopoly with imperfect substitutes. Yet, the monotonicity properties and efficiency results do not generally hold in other games. The next section gives an example: the cournot duopoly with imperfect substitutes. ■

7 Cournot duopoly

Let us now study an example to see what happens when supermodularity of the underlying is violated. We study a cournot dupoly where the strategy sets are $s_i \in [-10, ..., 0]$ and the underlying game payoff of player i reads

$$u_i(s_i, s_j) = \max\{-(\frac{19}{2} + \frac{1}{2}s_i + s_j)s_i, 0\} \tag{20}$$

It is easy to check that this game is symmteric and satisfies properties {1} and {2}, but not {3}: increasing player $i's$ action by one unit from s_i increases the payoff of the opponent:

$$u_i(s_i, s_j + 1) - u_i(s_i, s_j) = -s_i > 0 \tag{21}$$

and $\delta = -1$, $\sigma = 0$, and $\phi = -1$.

Notice that despite the negative strategies, this is indeed a game equivalent to a cournot duopoly with imperfect substitutes[20]. Vives (1989) shows that it can be transformed into an equivalent game which is supermodular by setting $\widetilde{s}_2 = -s_2$. Such a transformation would yield $\phi = 1 > 0$ and $\delta + 2\phi + \sigma = 1$. However, then both payoffs are not increasing in the action of the opponent.

Condition (18) requires that $i's$ marginal payoff, $-10 - s_j - s_i$, is non-positive if s is agreeable for i. Thus, an agreeable outcome satisfies $m \in \{m|10 + s_j + s_j \geq 0,\ i = 1, 2\}$.

Notice, that player i's underlying game best reply to s_j is

$$BR_i(s_j) = -10 - s_j \tag{22}$$

Thus the unique underlying game equilibrium is $s_1 = -5 = s_2$ which gives payoff $u_i^* = u_i(5, 5) = 10$ to both players. At this equilibrium, the benefit from breaching is exactly zero, $b(5, 5, 4) = 0$ as required in theorem 7.

Let's suppose that the guilt cost is as in (3). This guilt cost is supermodular and convex in u_i as required in proposition 6. By proposition 5 non-positive marginal incentive to breach is necessary and sufficient for incentive compatibility. Each player wants to deviate downwards. The marginal incentive to breach writes

$$10 + s_j + s_i + \theta_i[u_i(s) - 10]s_j \tag{23}$$

This is increasing in player's own action but the effect of opponent's action is ambiguous as opposed to proposition 6 which assumes that the game is supermodular.

Since $\delta = -1$, $\sigma = 0$ increasing player's agreed action increases player's marginal benefit from breaching and leaves the marginal harm unaffected. Wihtin the agreeable outcome set, the agreed payoff effects are as before: thus, agreed payoff decreases in own action. To sum up, the marginal incentive to breach is indeed increasing in player's own action.

Yet, now since the game is submodular, $\phi = -1$, rather than supermodular, increasing opponent's action decreases the marginal harm on the opponent and decreases player's marginal benefit from breaching. Agreed payoff increases in

[20] In an equivalent game, $\widetilde{s}_i \in [0, 10]$ and $\widetilde{u}_i(\widetilde{s}_i, \widetilde{s}_j) = \max\{(\frac{19}{2} - \frac{1}{2}\widetilde{s}_i - \widetilde{s}_j)\widetilde{s}_i, 0\}$ where $\widetilde{s}_i = -s_i$. The transformation is done in order to satisfy assumption {1}. Both the transformation and the orginal game are supermodular. The payoffs are chosen to make the best reply mapping simple. The analysis applies more generally.

own action, as before. The agreed payoff effect and the other two effects now run counter to each other. Thus, the effect on opponent's incentive to breach is ambiguous: the monotonicity of the marginal incentive to breach in agreed actions (proposition 6) is lost.

Now, let us move on and consider theorem 7 which studies whether efficient agreements can be made if any. Figure 3 studies the positive quantity equivalent of the game[21]. There, we suppose that the proneness to guilt is $\theta_i = \frac{1}{7}$ for both players. The outcomes marked with a plus sign are agreeable for player 1 and outcomes marked with a cross are agreeable for player 2. Thus, the outcomes marked with an asterisk belong to the agreeable outcome set. There are two symmetric outcomes in this set: the equilibrium $(5, 5)$ and the outcome where actions are decreased by one from the equilibrium, $(4, 4)$. Yet, the efficient symmetric outcome $(3, 3)$ (marked with a circle) does not belong to the agreeable outcome set.

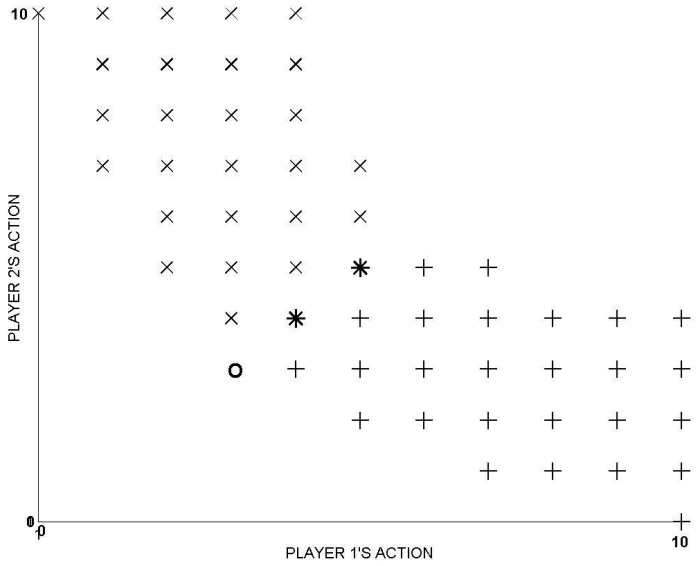

Figure 3: The agreeable set in the cournot duopoly.

To see that $(3, 3)$ is efficient, maximise

$$\max_{\sigma}\{-(\frac{19}{2} + \frac{3}{2}\sigma)\sigma\} \tag{24}$$

This is indeed concave in σ. Looking at first order effects, a unit increase in both actions increases the expression in the brackets if and only if $\sigma \leq -\frac{11}{3}$. The

[21] The relevant figure for the negative quantity game studied analytically is the projection of figure 3 through the orgin to the negative quadrant.

agreed payoffs for the symmetric outcomes corresponding to the nearest two integers of $\sigma = -\frac{11}{3}$are $u(-3,-3) = 15$ and $u(-4,-4) = 14$. Thus $s = (-3,-3)$ is efficient.

The underlying game equilibrium $(-5,-5)$ is agreeable by proposition 2. To see that $(-4,-4)$ is agreeable, we check that the marginal incentive to breach is negative, $10 - 4 - 4 - \frac{4}{7}[14 - 10] < 0$. Thus $s = (-4,-4)$ is agreeable. For $s = (-3 - 3)$, the marginal incentive to breach reads $10 - 3 - 3 - \frac{3}{7}[15 - 10] = \frac{13}{7} > 0$ and thus, for $\theta_i = \frac{1}{7}$ $i = 1, 2$, players can agree on $s = (-4,-4)$ but not on $s = (-3,-3)$.

Even if the outcome where actions are increased by one from the symmetric equilibrium belongs to the agreeable outcome set, the efficient symmetric agreement does not belong to that set! The property of theorem 7 does not hold in the cournot dupoly. Yet, the cournot duopoly satisfies all other assumptions of that proposition but supermodularity of the underlying game.

This is because marginal symmteric changes of both actions (i) increase the marginal harm by $-\delta - \phi$ where both terms are strictly positive, decrease the marginal harm by $\sigma + \phi$ where the first term is zero and the second term is strictly negative, and (iii) and change the marginal effect of an increasing agreed payoff by $\delta + 2\phi + \delta < 0$. The negative marginal effect on the marginal incentive to breach is vanishing but the positive marginal effects are constant and thus getting relatively stronger as agreed payoff is increased by symmteric changes of both actions. Thus, even if there is a non-equilibrium action where guilt overtakes the underlying game incentive to breach, the incentives to respect a more efficient action profile are smaller. Consequently, we also lose any efficiency property akin to that in theorem 7.

8 Discussion

We have shown that the interaction and equilibrium outcomes may change rather substantially due to preplay negotiation when players feel guilt if they unilaterally breach an oral agreement. Preplay negotiation may alter the payoffs of the underlying game and transform dilemmas into coordination games the equilibrium selection problem of which is dissolved in a manner we are familiar with from the work of Farrell (1987).

The theory presented is in line with the results from public goods game experiments without tresholds where communication significantly increases contribution levels (Ledyard, 1995). Our theory predicts that in the public good game, reaching the symmetric efficient agreement is more likely to occur than reaching a symmetric inefficient non-equilibrium agreement. Also, conditional on reaching such an agreement, breaching is less likely to occur than breaching conditional on reaching an inefficient agreement. Furthermore, increasing the marginal per capita return facilitates agreeing on non underlying game equilibrium play. In the prisoner's dilemma, increasing the benefit of defection and decreasing the harm on the opponent when the opponent cooperates, will make agreeing on cooperation less likely to occur.

Furthermore, for results that apply for a broader class of games, Nash equlibria are always agreeable. In games with concave payoffs in each action, broadly, cheking marginal incentive to breach from the agreement is a necessary and a sufficient condition to check whether an agreement is agreeable. In supermodular games, the marginal incentive to breach is monotone in each agreed action when the guilt cost is supermodular in the agreed payoff and in the harm. Furthermore in symmetric supermodular games, a non-underying game equilibrium is agreeable if and only if an efficient outcome is agreeable. These games include the public good game, moral hazard in teams, and some bertrand dupolies with imperfect subsitutes. In other games, both the monotonicity of the marginal incentive to breach and the ability to agree efficiently even if some non-equilibria are agreeable, do not hold.

When incorporating the guilt cost, it is important to carefully consider what factors affect guilt that is felt when acting counter to an ideal of keeping an agreement. Our assumed characteristics are well established in psychological research.

Analogous results as in this paper would obtain, if we suppose that players have zero proneness to guilt and they informally agree on a stationary action profile in an infinitely repeated analog of the underlying game with continuous time. The punishment paths are not negotiated, however, but they are exogenously determined (in a commonly known social contract, for instance). If the agreement is breached, it takes some time to detect breaching and when detected players revert to mutual minmax strategies for a length of time that depends on the agreed payoff and the harm inflicted on the other. As stated in the introduction, the origin of guilt, according to psychologists, resides in such close communal relationships where the prevailing social contract gets internalised.[22]

This paper has not analysed the effects of negotitation protocol on the agreement and the outcome. We intend to study this in a follow-up paper which also presents expermental evidence of such effects. Any bargaining protocol has an preplay negotiation analog, Nash demand negotiation or the ultimatum negotiation being the simplest examples. Furthermore, there are focal non-bargaining protocols that seem particularly relevant for preplay negotiation with guilt-prone players. A globally applied and a simple preplay negotiation institution is promising where players simultaneously or sequentially make a promise of which action they will choose in the underlying game.

As the negotiation protocol is unanalysed, the prediction of the model is generally only setwise and thus not very sharp. Combining the model with a negotiation protocol would make the prediction sharper. For instance, the prediction of any bargaining game would coincide with the prediction of that bargaining game when the only available outcomes are the agreeable outcomes and the available payoffs are those associated with the agreeable outcomes. The disagreement point or the outside option would coincide with the worst underlying game Nash equilibria and the associated payoffs.

Another dimension for future research is the relaxation of the assumption

[22] See appendix for further details.

of complete information on proneness to guilt types. The choice of an optimal agreement when information is private seems to involve trading off own agreed payoff and the probability that the opponent breaches the agreement[23]. Yet, to find conditions that guarantee that an efficient action profile is proposed or that such a local optimum is a global optimum may turn out difficult.

On the other hand, a dynamic setup of incomplete information on proneness to guilt would allow players to build up reputations. First, it may be optimal for types with high proneness to guilt to build up a reputation for a lower proneness to guilt so that they are proposed higher shares of the surplus in the future. Second, types with a low proneness to guilt are willing to build up a reputation for a higher proneness to guilt in order to be able to reach agreements with a larger fraction of types.

9 Appendix

9.1 Proof of lemma 1

Proof. We use the proof by contradiction. Assume that $m \in BR_i(m_j)$ and assume to the contrary that a deviation from m is profitable when m is agreed. That is $U_i(s_i, m_j, m) = u_i(s_i, m_j) - \theta_i g(u_i(m), h_j(m, s_i)) > u_i(m) = U(m, m)$ where the first equality follows from (BD) and the fact that $s_i \neq m_i$ and $s_j = m_j$. But by assumption $g(u_i(m), h_j(m, s_i)) \geq 0$. Hence, $u_i(s_i, m_j) \geq u_i(s_i, m_j) - \theta_i \gamma(m, s_i, m_j) > u_i(m)$. This is a contradiction, because m is an underlying game best reply. Hence, no deviation is profitable. Thus, $m \in A_i(\Gamma, \theta_i)$ iff $u_i(m) \geq u_i^*$. ■

9.2 Proof of proposition 2

Proof. Since m is an equilibrium $u_i(m) \geq u_i^*$ for $i = 1, 2$. Since m is an equilibrium, ,$m_i \in BR_i(m_j)$ for $i = 1, 2$. Then, by lemma (1), $m \in A_i(\Gamma, \theta_i)$ for $i = 1, 2$ and $m \in A(\Gamma, \theta)$. ■

9.3 Proof of lemma 2

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

9.4 Proof of lemma 3

Proof. Conditions (AC) and (EF) imply that $g_i(u_i(m), h_j(m, s_i)) = 0$. But,

[23] Notice yet, that if the information on proneness to guilt is private, signalling is not an issue: 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 and, thus, her opponent knows that she will not suffer guilt and thus only underlying game best replies are agreeable.

$h_j(m, s_i) \doteq u_j(m) - u_j(m_j, s_i) \leq 0$. Thus the incentive compatibility of player i is violated. ■

9.5 Proof of proposition 3

Proof. Assume that m satisfies $u_i(m) > \min_{s^* \in NE(\Gamma)}\{u_i(s^*)\}$ for $i = 1, 2$. By lemma 3, if for some i and some s_i $u_i(m) \geq u_i(s_i, m_j)$ or $u_j(m_j, s_i) < u_j(m)$ does not hold, then $m \notin \lim_{\theta_1 \to \infty, \theta_2 \to \infty} A(\Gamma, \theta)$. If for all i for all s_i $u_i(m) \geq u_i(s_i, m_j)$ or $u_j(m_j, s_i) < u_j(m)$, assume to the contrary that the agreement does not belong to $\lim_{\theta_1 \to \infty, \theta_2 \to \infty} A(\Gamma, \theta)$. Then there is an i and a s_i for which $u_i(m) \geq u_i(s_i, m_j)$ or $u_j(m_j, s_i) < u_j(m)$. The incentive compatibility condition is captured by $\lim_{\theta_i \to \infty} \theta_i g(u_i(m), h_j(m, s_i)) \geq u_i(s_i, m_j) - u_i(m)$ where, by assumption, the left-hand side is positive infinite if $h_j(m, s_i) > 0$ and 0 otherwise. If $u_i(m) \geq u_i(s_i, m_j)$ holds, then the right hand side is negative and we reach a contradiction. If $u_j(m_j, s_i) < u_j(m)$ then $h_j(m, s_i) > 0$ and the left-hand side is positive infinite whereas the right hand side is finite. We reach a contradiction.Proof of ■

9.6 Proof of proposition 5

Lemma 5 *Let Γ be finite. Let $m_i \neq [\underline{s}_i, \overline{s}_i]$. Let {1} , {2} and {4} hold. Then an action profile is agreeable if and only if the marginal incentive to breach is non-positive.*

Proof. (Necessity) Let $\beta_i(m, \theta_i) > 0$. Then by definition, it is optimal for player i to breach.

(Sufficiency). Suppose now that deviation to s_i' is profitable and to the contrary that $\beta_i(m, \theta_i) \leq 0$. Suppose that $u_i(m_i + 1, m_j) - u_i(m_i, m_j) > 0$. But assumption {EF} leads to a contradiction, since $h_j < 0$ and thus guilt cost is zero.

Suppose that $u_i(m_i - 1, m_j) - u_i(m_i, m_j) \geq 0$. By assumption $\beta_i \leq 0$ and thus

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

By assumption {1} the harm is increasing 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. Thus the cost is convex in s_i. On the other hand, by assumption {2} the payoff u_i is concave in s_i, the benefit from breaching $u_i(s_i, m_j) - u_i(m_i, m_j)$ is concave in s_i. Thus if $\beta_i(m, \theta_i) \leq 0$ then $B(m, s; \theta_i) \leq 0$ for all $s_i < m_i$. We have a contradiction. ■

9.6.1 Proof of the proposition

Proof. The result follows from lemma 5 and the monotonicity of the guilt cost in $u_i(m)$ and the assumption that $g(u_i(m), h_j) = 0$ for $u_i(m) = u_i^*$. Hence, if the incentive compatibility constraint is satisfied, then the individual rationality

is satisfied as well. [NOTICE RESERVATION, holds only if $b_i > 0$ for $i = 1, 2$ since by setting $b_i(m) = 0$ there could be $u_i(m) < u_i^*$ and m agreeable if $g_j(u_j, h_i) > 0$] ■

9.7 Proof of proposition 6

Proof. Since u_i is increasing in s_j, by lemma 3, we need u_i to be decreasing in s_i for (s_i, s_j) to be agreeable. Then, the marginal incentive to breach writes

$$\beta_i(m_i, m_j) = b_i(m_i, m_j, m_i - 1) - \theta_i g(u_i(m_i, m_j), h_j(m_i, m_j, m_i - 1))$$

But $b_i(m_i, m_j, m_i - 1)$ is increasing in m_i since $u_i(m_i, m_j) - u_i(m_i - 1, m_j)$ is negative and u_i is concave in m_i. Also, $u_i(m_i, m_j) - u_i(m_i - 1, m_j) < 0$ implies that $u_i(m)$ decreases in m_i. Furthermore, $h_j(m_i, m_j, m_i - 1)$ is decreasing in m_i since $u_j(m_j, m_i) - u_j(m_j, m_i - 1)$ is positive and u_j is concave in m_i. But g is increasing in both arguments. Thus, $\beta_i(m_i, m_j)$ is indeed increasing in m_i.

On the other hand, $b_i(m_i, m_j, m_i - 1)$ is decreasing in m_j since u_i is supermodular. Also, u_i is increasing in m_j by assumption. Finally, $h_j(m_i, m_j, m_i - 1)$ is increasing in m_j since u_j is supermodular. ■

9.8 Proof of theorem 7

We can use shorthands $b(s), u(s)$ and $h(s)$ for $b(s_i, s_j, s_i - 1), u(s_i, s_j)$ and $h(s_i, s_j, s_i - 1)$ respectively. Let us denote $b(s_i + 1, s_j + 1)$ by $b(s + 1)$ and $u(s + 1) \doteq u(s_i + 1, s_j + 1)$ and $h(s + 1) \doteq h(s_i + 1, s_j + 1)$.

Also $s' > s$ means that there is integer $k > 0$ such that $s'_i > s_i + k$ for $i = 1, 2$.

Lemma 6 *Then* $b(s + 1) - b(s) = -\delta - \phi$

Proof. On the one hand,

$$u_i(s, s + 1) - u_i(s + 1, s + 1) - [u_i(s, s) - u_i(s + 1, s)] = -\phi$$

and on the other hand

$$u_i(s, s) - u_i(s + 1, s) - [u_i(s - 1, s) - u_i(s, s)] = -\delta$$

Thus

$$\begin{aligned} b(s + 1) - b(s) &= u_i(s, s + 1) - u_i(s + 1, s + 1) - [u_i(s - 1, s) - u_i(s, s)] \\ &= u_i(s, s + 1) - u_i(s + 1, s + 1) - [u_i(s, s) - u_i(s + 1, s)] \\ &\quad + u_i(s, s) - u_i(s + 1, s) - [u_i(s - 1, s) - u_i(s, s)] \\ &= -\delta - \phi \end{aligned}$$

■

Lemma 7 *If $\delta \neq \phi$ then there is at most one equilibrium s^* where $b_i(s^*) = 0$ for $i = 1, 2$*

If $-\delta > \phi$ then $(s_i, s_j\ ) \in M_F$ implies $s_i > s^*$ for $i = 1, 2$

If $-\delta < \phi$ then $(s_i, s_j\ ) \in M_F$ implies $s_i < s^*$ for $i = 1, 2$

Proof. As a mapping from S_2 to S_1 the best reply curve of player one, $BR_1^{-1}(m_1)$, has slope $-\frac{\delta}{\phi}$ and that of player two, $BR_2(m_1)$, has slope $-\frac{\phi}{\delta}$ which are positive constants. The crossing point of the BR curves is a unique symmetric equilibrium, $s^* = (s_1^*, s_2^*)$. M_F implies that $\frac{\partial u_i(s)}{\partial s_i} \leq 0$ for $i = 1, 2$. For player two this is true above $BR_2(m_1)$ and for player one this is true to the right of $BR_1^{-1}(m_1)$. Thus the claim. ■

Lemma 8 $h(s+1) - h(s) = \sigma + \phi$

$$\begin{aligned} & h(s+1) - h(s) \\ = \ & u(s+1, s+1) - u(s+1, s) - u(s, s) + u(s, s-1) \\ = \ & u(s+1, s+1) - 2u(s+1, s) + u(s+1, s-1) \\ & + u(s+1, s) + u(s, s-1) - u(s+1, s-1) - u(s, s) \\ = \ & \sigma + \phi \end{aligned}$$

Lemma 9

$$\begin{aligned} & u(s'') - u(s') \\ = \ & -\frac{(s'' - s' - 1)(s" - s' - 2)}{2}(\sigma + 2\phi + \delta) + (s" - s')(u(s") - u(s" - 1)) \end{aligned}$$

Proof. First, notice that (proof of lemma 13)

$$\begin{aligned} & u_i(s+2, s+2) - u_i(s+1, s+1) - [u_i(s+1, s+1) - u_i(s, s)] \\ = \ & u_i(s, s) - u_i(s+1, s) - u_i(s+1, s) + u_i(s+2, s) \\ & + u_i(s+2, s+2) - u_i(s+2, s+1) - u_i(s+2, s+1) + u_i(s+2, s) \\ & + u_i(s+2, s+1) - u_i(s+2, s) - u_i(s+1, s+1) + u_i(s+1, s) \\ & + u_i(s+2, s+1) - u_i(s+2, s) - u_i(s+1, s+1) + u_i(s+1, s) \\ = \ & \sigma + \delta + 2\phi \end{aligned}$$

Furthermore, let $s'' - s' = 2n \; n \in N$

$$\begin{aligned}
& u(s") - u(s') \\
= & \sum_{k=0}^{s"-s'-1} u(s'+k+1) - u(s'+k) \\
= & \sum_{k=0}^{s"-s'-2} -u(s'+k) + 2u(s'+k+1) - u(s'+k+2) \\
& - \sum_{k=0}^{s"-s'-2} u(s'+k+1) + \sum_{k=2}^{s"-s'-2} u(s'+k) + 2u(s") \\
= & -(s" - s' - 2)(\sigma + 2\phi + \delta) + u(s") - u(s'+1) + u(s") - u(s"-1)
\end{aligned}$$

Thus,

$$\begin{aligned}
u(s") - u(s') = & -(s" - s' - 1)(\frac{(s" - s' - 2)}{2})(\sigma + 2\phi + \delta) \\
& +(s" - s')(u(s") - u(s" - 1))
\end{aligned}$$

■

Lemma 10 *Let the game be symmetric. If s^c maximises $max_{k \in Z} u(s+k)$ where $s_i = s_j$ (along the diagonal) then there is no s' such that $u_i(s') > u_i(s^c)$ for $i = 1, 2$.*

Such s^c exists.

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

Since S is finite and $u(s+k)$ is defined for all $k \in Z$, there must be a k where that maximises $u(s+k)$ with $s_i = s_j$. ■

Lemma 11 *When not in M_F, there is i such that $u_i(s)$ is increasing in symmetric unit increments of both actions*

Proof. Not in M_F , there is a player i such that $u_i(s+1) - u_i(s) = [u_i(s+1, s+1) - u_i(s, s+1)] + [u_i(s, s+1) - u_i(s, s)] > 0$. Thus not within M_F, $u_i(s)$ is increasing in s (in symmetric unit increments of both actions). ■

Lemma 12 *Let $b(s^*) = 0$. Suppose that $\sigma + 2\phi + \delta \geq 0$ implies*

$$\begin{aligned}
& \frac{(s^* - \underline{s} - 2)(s^* - \underline{s} - 1)}{2(s^* - \underline{s})}(\sigma + 2\phi + \delta) \qquad (25) \\
\leq \; & [u(s^* - 1, s^*) - u(s^* - 1, s^* - 1)]
\end{aligned}$$

Then no $s' < s^*$ satisfies $u(s') > u(s^*)$

Proof. First, we show that this holds for $s^* - 1$:

$$\begin{aligned} u(s^*) - u(s^* - 1) &= u(s^*, s^*) - u(s^* - 1, s^*) + [u(s^* - 1), u(s^*)] - u(s^* - 1, s^* - 1)] \\ &= -b(s^*) + [u(s^* - 1, s^*)] - u(s^* - 1, s^* - 1)] \\ &> 0 \end{aligned}$$

where the last row follows from the fact that the payoff is increasing in opponent's action.

Using lemma 9, we can write $u(s^*) - u(s)$ for $s < s^*$ as follows

$$\begin{aligned} & u(s^*) - u(s) \qquad (26) \\ = & -\frac{(s^* - s - 1)(s^* - s - 2)}{2}(\sigma + 2\phi + \delta) \\ & +(s^* - s)(u(s^*) - u(s^* - 1)) \\ = & -\frac{(s^* - s - 1)(s^* - s - 2)}{2}(\sigma + 2\phi + \delta) \\ & +(s^* - s)[-b(s^*) + u(s^* - 1, s^*) - u(s^* - 1, s^* - 1)] \end{aligned}$$

The second term is positive since the payoff is increasing in the opponent's action and since $b(s^*) = 0$. There are two subcases to consider: i) $\sigma + 2\phi + \delta \leq 0$ and ii) $\sigma + 2\phi + \delta > 0$. In case, i) the first term of (26) is positive. In ii) since

$$\begin{aligned} & \frac{(s^* - \underline{s} - 2)(s^* - \underline{s} - 1)}{2(s^* - \underline{s})}(\sigma + 2\phi + \delta) \\ \leq & [u(s^* - 1, s^*) - u(s^* - 1, s^* - 1)] \end{aligned}$$

where $\underline{s}$ is the smallest feasible action, (26) is non-negative. Thus the claim holds. ■

Lemma 13 *Let y be convex and supermodular. Then $y(x + 2, z + 2) - 2y(x + 1, z + 1) + y(x, z) \geq 0$*

Proof. Let y be convex and supermodular. Then

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

The first effect on the RHS 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 supermodularity effect. ■

Proposition 8 *Let $\phi \geq -\sigma$ and $-\delta > \phi$. Suppose that {1}, {2}, and {3} hold. Furthermore, let g given by 3. If $(\underline{s}, \underline{s})$ and $(\overline{s}, \overline{s})$ are agreeable and $b(\underline{s}-1, \underline{s}-2) \geq 0$, $u_i(\underline{s}) \geq u_i^*$, $\underline{s}-1 \notin A_i(\Gamma, \theta_i)$ and $u_i(s)$ is increasing in s for $\underline{s} \leq s \leq \overline{s}$ and for $i = 1, 2$ then any (s, s) such that $\underline{s} < s < \overline{s}$ is agreeable.*

Proof. By lemma 6, $b(s)$ is increasing in s. Since $\phi \geq -\sigma$, by lemma 8 $h(s)$ is increasing in s. By assumption $u(s)$ is increasing in s within $[\underline{s}, \overline{s}]$.then

$$g(u(s), h(s)) \leq g(u(s+1), h(s+1))$$

since g is weakly increasing in u and in h.

By lemma 6, $b(s)$ is linear in s. Since $\underline{s} - 1 \in M_F$ and $\underline{s} - 1 \notin A_i(\Gamma, \theta_i)$, we have $b(\underline{s}-1) > g(u(\underline{s}-1), h(\underline{s}-1))$. Then we need that

$$\begin{aligned} & g(u(\underline{s}), h(\underline{s})) - g(u(\underline{s}-1), h(\underline{s}-1)) && (27) \\ \geq \quad & b(\underline{s}) - b(\underline{s}-1) && (28) \end{aligned}$$

for $\underline{s}$ to be agreeable.

Suppose now that there is s' such that $\underline{s} < s' < \overline{s}$ and (s', s') is not agreeable but $s'-1$ is agreeable. (Clearly, such a non-agreeable s' must exist if any because otherwise $\underline{s}$ was not agreeable.). Then because $b(\underline{s}-1) \geq 0$, $b(\overline{s}) \geq 0$ also $b(s'), b(s'-1) \geq 0$ and

$$\begin{aligned} g(u(\underline{s}), h(\underline{s})) - g(u(\underline{s}-1), h(\underline{s}-1)) \quad & \geq \quad b(\underline{s}) - b(\underline{s}-1) = b(s') - b(s'-1) \\ & \geq \quad g(u(s'), h(s')) - g(u(s'-1), h(s'-1)) \end{aligned}$$

If $u(s)$ is convex in s, then $g(u(s), h(\widehat{s}))$ is convex in s since g is linear in u. Similarly, $g(u(\widehat{s}), h(s))$ is convex in s since g is linear in h for $h \geq 0$. Also by lemma 8, $h(s)$ is increasing in s and by assumption $u(s)$ is increasing in s. Thus,

$$\begin{aligned} & g(u(\underline{s}), h(\underline{s})) - g(u(\underline{s}-1), h(\underline{s}-1)) \\ \leq \quad & g(u(s'), h(s')) - g(u(s'-1), h(s'-1)) \end{aligned}$$

which is a contradiction.

On the other hand, if $u(s)$ is not convex in s then it must be strictly concave. But then $g(u(s), h(\widehat{s}))$ is concave in s for $u(s) \geq u_i^*$ and $g(u(\widehat{s}), h(s))$ is concave in s since g is linear in h when $h \geq 0$. Also by lemma 8, $h(s)$ is increasing in s and by assumption $u(s)$ is increasing in s. Thus,

$$\begin{aligned} & g(u(\underline{s}), h(\underline{s})) - g(u(\underline{s}-1), h(\underline{s}-1)) \\ \leq \quad & g(u(s'), h(s')) - g(u(s'-1), h(s'-1)) \end{aligned}$$

which is a contradiction. ■

Proposition 9 ??*Let $\delta + 2\phi + \sigma > 0$. Let $(\underline{s}-1, \underline{s}-1) \in M_F$ and $(\underline{s}-1, \underline{s}-1) \notin A(\Gamma, \theta)$. Let $u(\underline{s}) - u(\underline{s}-1) \geq 0$. Suppose that {1}, {2}, and {3} hold. Furthermore, let g satisfy {4}, and {5}. If $(\underline{s}, \underline{s})$ such that $\underline{s}_i > s_i^*$ is agreeable then any (s, s) such that $\underline{s} < s$ is agreeable.*

Proof. $\delta + 2\phi + \sigma > 0$ implies that $-\delta < \phi$ or $-\sigma < \phi$. Suppose first that $-\delta < \phi$. But by lemma 7 $\underline{s}_i - 1 < s_i^*$ for $i = 1, 2$. Thus $\underline{s}_i \leq s_i^*$ and the claim holds trivially.

Let now, $\phi \leq -\delta$. Then $-\sigma < \phi$. The fact that $\underline{s} - 1$ is in M_F implies that $b(\underline{s} - 1, \underline{s} - 1, \underline{s} - 2) \geq 0$. By lemma 6, $b(s)$ is non-decreasing in s (symmteric unit increments). Also, since $\phi > -\sigma$, by lemma 8 h is linearly increasing in s and thus convex in s.

On the one hand, $u(\underline{s}+1) - u(\underline{s}) \geq u(\underline{s}) - u(\underline{s}-1) \geq 0$ since $\sigma + 2\phi + \delta > 0$. On the other hand, $u(\underline{s}) \geq u^*$ since $\underline{s}$ is agreeable.

Since $u(s)$ is convex in s, then $g(u(s), h(\widehat{s}))$ is convex in s since g is convex in u. Similarly, $g(u(\widehat{s}), h(s))$ is convex in s since g is convex in h for $h \geq 0$.

Also since $(\underline{s} - 1, \underline{s} - 1) \notin A(\Gamma, \theta)$ but $(\underline{s}, \underline{s}) \in A(\Gamma, \theta)$, by (27),

$$\begin{aligned} & b(\underline{s}) - b(\underline{s} - 1) \\ \leq \; & g(u(\underline{s}), h(\underline{s})) - g(u(\underline{s} - 1), h(\underline{s} - 1)) \end{aligned}$$

Thus, by lemma 13 and since g is supermodular in its arguments

$$\begin{aligned} 0 \; &\leq \; b(\underline{s} + 1) - b(\underline{s}) \\ &= \; -\delta - \phi \\ &= \; b(\underline{s}) - b(\underline{s} - 1) \\ &\leq \; g(u(\underline{s}), h(\underline{s})) - g(u(\underline{s} - 1), h(\underline{s} - 1)) \\ &\leq \; g(u(\underline{s} + 1), h(\underline{s} + 1)) - g(u(\underline{s}), h(\underline{s})) \end{aligned}$$

We can proceed by induction to show that for every $s > \underline{s}$, we have $u(s) > u(\underline{s})$ and that $b(s) - g(u(s), h(s)) \leq b(\underline{s}) - g(u(\underline{s}), h(\underline{s})) \leq 0$. Thus every, $s > \underline{s}$ such that $u(s) > u(\underline{s})$ is agreeable. ■

Corollary 1 *Suppose that $\sigma + 2\phi + \delta > 0$. Let Γ be symmteric. Suppose that {1}, {2}, and {3} hold. Furthermore, let g satisfy {4}, and {5}. Let s^* be the unique equilibrium of the game with $b(s^*, s^*, s^* - 1) = 0$. The an agreement in the core, s^c is agreeable if $s^* + 1$ is agreeable.*

By assumption $2\phi + \sigma + \delta > 0 = u(s + 2) - 2u(s + 1) + u(s)$. Thus, $u_i(s)$ is convex in s. Since $b(s^*, s^*, s^* - 1) = 0$, $s^* \in M_F$. By setting $\underline{s} = s^* + 1$, it follows from proposition **??** that $s \geq s^*$ is agreeable if $s^* + 1$ is agreeable.

It remains to show that among these agreements there is one in the core. Since $u(s^*) - u(s^* - 1) > 0$ and $2\phi + \sigma + \delta > 0$ the agreement with one of the actions at $\overline{s}$ pareto-dominates other action profiles reachable with symmteric increments from s^*. Denote this s^c. But we saw above that $u(s^c) \geq u(s^*)$. Notice that if $s^* + 1$ is agreeable, then $-\delta \geq \phi$ because otherwise only s^* is agreeable. But this implies that any $s_i < s_i^*$ is not in M_F. Thus by lemma 11, there is i such that $u_i(s) < u_i(s^*)$ along the diagonal. Finally, lemma 10 shows that no asymmetric agreeable action profile pareto-dominates s^c.

Theorem 10 *Suppose that $\sigma + 2\phi + \delta \geq 0$. Suppose that the game is symmetric and that {1}, {2}, and {3} hold. Furthermore, let g satisfy {4}, and {5}.*

Let s^* be the unique equilibrium of the game and $b_i(s^*, s^*, s^* - 1) = 0$. Then an agreement in the core, $s_i^c \geq s^*$ $i = 1, 2$ is agreeable if and only if some $s_i > s^*$, $i = 1, 2$ and not on the boundary of M_F is agreeable.

$\sigma + 2\phi + \delta \geq 0$ implies that either $\delta + \phi \geq 0$ or $\sigma + \phi \geq 0$.

If $\phi > -\delta$ then by lemma 7 $s \in M_F$ implies that $s_i < s^*$ $i = 1, 2$. Thus the claim holds trivially.

If $\delta + \phi = 0$ then $b(s+1) - b(s) = 0$ and either there are multiple equilibria or in the unique equilibrium there is i such that $b_i(s^*, s^*, s^* - 1) \neq 0$ both contrary to our assumptions.

If $\phi < -\delta$ then, $\sigma + \phi \geq 0$.

For any $s \notin M_F$, there is i such that $u_i(s) - u_i(s-1) > 0$. But since $\sigma + 2\phi + \delta \geq 0$, then for any $s' > s$ $u_i(s') - u_i(s' - 1) > 0$.

Suppose that s is agreeable and it is not on the boundary of M_F. Since the best reply curves have a slope less steep than one, $s - 1$ is in M_F. [Off diagonal argument and on diagonal argument]. Then any $s' > s$ is agreeable by proposition **??**.

To complete the proof, we need to show that for any s there is $s' > s$ that is efficient. We will first show that any s where there is player i such that $s_i = \overline{s}_i$ is efficient. We first show this for $(\overline{s}_1, \overline{s}_2)$.

When not in M_F, the payoff is decreasing for at least one of the players. Thus along the diagonal (where both choose the same action), $u(s) < u(s^*)$ for $s < s^*$. On the other hand, $u(s^*) - u(s^* - 1) = -b(s^*) - [u(s^* - 1, s^*) - u_i(s^* - 1, s^* - 1) > 0$ and since $\sigma + 2\phi + \delta \geq 0$, $u(s) > u(s^*)$ for all $s > s^*$ along the diagonal. Thus $(\overline{s}_1, \overline{s}_2)$ is efficient by lemma 10.

Outside M_F, there is a player the action of which can be increased thereby increasing the payoff of both. Thus, no profile not in M_F is efficient. Where $b_i(s) = 0$, the slope of the indifference curve of player i is zero and the slope of the indifference curve of j is greater than zero. Thereby, a symmteric increment of both actions improves the payoff of both. Since $\sigma + 2\phi + \delta \geq 0$, such increments improve payoff of both for all $s' > s$ and since the best reply curves are less steep than one, then such increments keep the action profile within M_F. Thus only action profiles where there is i such that $s_i = \overline{s}_i$ can be efficient. On the other hand, such profiles are not pareto-ranked when within M_F. Thus, all such action profiles are efficient. This completes the proof.

9.9 Repeated Games

Analogous results as in this paper would obtain, if we suppose that players have zero proneness to guilt and they informally agree on a stationary action profile in an infinitely repeated game with continuous time. The punishment paths are not negotiated, however, but they are exogenously determined (in a commonly known social contract, for instance). If the agreement is breached players revert to mutual minmax strategies and the punishment phase lasts for time interval $k(.)$ and the length of the punishment depends on the agreed payoff and the harm.

If such punishment paths indeed reflect a common sense of justice prevailing in society, then, in one-shot games, the guilt cost might serve as an internalised punishment that reflects society's sense of justice. Psychologists such as Clark and Mills (1979) argue for such origins of guilt.

It is easily verified that to make the incentives to breach exactly identical as in the case of guilt, we must make the following assumptions

- discount rates are equal $\rho_i = 1$ for $i = 1, 2$
- It takes time $w = -\ln(\frac{1}{2})$ to observe that opponent is breaching.
- the punishment function $k(h_j(m, a_i), u_i(m), u_i^P)$ takes the following form

$$k(h_j(m, a_i), u_i(m), u_i^P) = \lim_{\varepsilon \to 0} -\ln(\max\{\varepsilon, 1 - \theta \frac{g(u_i(m), h_j(m, s_i))}{u_i(m) - u_i^P}\})$$

(with u_i^P the mutual minmax payoff for player i). Yet, this formulation, implies that an infinitely long punishment follows a breaching where $u_i(m) - u_i^P \leq \theta g(u_i(m), h_j(m, s_i))$.

9.10 Social Contract Literature

In the game theoretic social contract literature, Harsanyi (1977) and later Binmore (1994), present models where, in addition to their individual preferences, players have empathetic preferences which are used to derive a perception of fairness.

The fairness preferences are derived from weighting of the empathetic preferences in an impartial orginal position where the player thinks it is equally likely that one ends up playing one's own role or that of the opponent. Empathetic preferences are defined over the set $S \times \{1, 2\}$ where S is the set of outcomes of play and $\{1, 2\}$ is the set of possible roles. Player has an ordering over the outcomes of the game faced either as oneself or as the opponent. Full empathy says that the ordering of S coincides with that of $u_i(s)$ for each i. This leads to a utility function which is a weighted sum of the preferences of the two players.

If the player uses his fairness preferences when playing the game after communication and considering a deviation that decreases the opponent's payoff, the guilt cost takes the form of example 3. The formulation $U_i(m, s) = u_i(s) + \theta_i u_i(m) h_j(m, s)$ is reached by letting the weight depend on the agreement m. The implication is thus an truncated additive social welfare function where the concern for the opponent depends on how nicely one is treated in the communication and how guilt averse (empathetic) one is.

10 References

References

[1] Andreoni, J. (2005): Trust, Reciprocity, and Contract Enforcement: Experiments on Satisfaction Guaranteed. University of Wisconsin. Mimeo.

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

[3] 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.

[4] Baumeister, R.F., Stillwell, A.M., & Heatherton, T.F. (1994): Guilt: An interpersonal approach. Psychological Bulletin. 115(2):243-267

[5] Baumeister, R.F., Stillwell, A.M., & Heatherton, T.F. (1995): Guilt as interpersonal phenomenon: Two studies using autobiographical narratives. In Self concious 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] Binmore, K. (1994): Game Theory and The Social Contract: Playing Fair. MIT Press. Cambridge, MA.

[8] Bornstein, G. (1992). Group decision and individual choice in intergroup competition for public goods. In W. Leibrand, D. Messick, & H. Wilke (Eds.), Social dilemmas: Theoretical issues and research findings (pp. 247-263). Oxford, UK: Pergamon Press.

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

[10] Charness, G.; Dufwenberg M. (2003): Promises, Promises. IUI Stockholm Working Paper.

[11] Crawford, V.; Sobel J. (1982): Strategic Information Transmission. Econometrica 50: 579-594.

[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 exhange and communal relationships. Journal of personality and social psychology 37: 12-24.

[14] Cox, J.; Friedman D. (2002). A Tractable Model of Reciprocity and Fairness. Mimeo.

[15] Dawes, Orbell, van de Kragt (1990): The Limits of Multilateral Promising. Ethics, 100: 616-627.

[16] 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.

[17] Duffy, J.; Feltowich N. (2004): Words, Deeds and Lies: Strategic Behavior in Games with Multiple Signals. Mimeo.

[18] Dufwenberg Martin (2002): Marital investments, time consistency and emotions. Journal of Economic Behavior & Organization Vol. 48 57-69

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

[20] Farrell. J. (1987): Cheap Talk, Coordination, and Entry. Rand Journal of Economics 18 (1): 34-39.

[21] Farrell, J.; Rabin M.(1996): Cheap Talk. Journal of Economic Perspectives. 10/3:103-118.

[22] Fehr, E.; Schmidt K. (1999): A Theory of Fairness, Competition and Cooperation. Quarterly Journal of Economics.

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

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

[25] Gneezy, U. (2005): Deception: The Role of Consequences. American Economic Review. Forthcoming.

[26] Harsanyi, J. (1977) Rational Behavior and Bargaining Equilibrium in Games and Social Situations. Cambridge, England: Cambridge University Press.

[27] 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.

[28] Holmström, B. (1982): Moral Hazard in Teams. Bell Journal of Economics 13 (2): 324-340.

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

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

[31] Isaac, M.; Walker J. (1988): communication and free-riding behavior: the volontary contribution mechanism. Economic Inquiry. 26(2): 586-608.

[32] 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

[33] Loomis, J. (1959): communication: The Development of Trust and Cooperative Behavior. Human Relations 12(3): 305-315.

[34] Millar, K.U., Tesser A. (1988): Deceptive behavior in social relationships: a consequence of violated expectations. Journal of psychology 122: 263-273.

[35] Nash, J. (1953): Two-person cooperative games. Econometrica, 21:128–140.

[36] Okel, E.; Mosher D. (1968): Changes in affective states as a function of guilt over aggressive behaviour. Journal of consulting and clinical psychology 32: 265-270.

[37] Pavitt, C.; Shankar, A. (2002): Resource and Public Good Dilemmas: A New Issue for Communication Research. The Review of Communication. 2.3: 251-272.

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

[39] Radlow, R; Weidner, M. (1966): Unforced commitments in 'cooperative' and 'non-cooperative' non-constant-sum games. Journal of Conflict Resolution 10:497-505.

[40] Vives, X. (1990): Nash Equilibrium with Strategic Complementarities. Journal of Mathematical Economics, 19, 305-321

[41] Smith A. (1759): The Theory of Moral Sentiments. Reprinted in (2002). Ed. Knud Haakonsen. Cambridge University Press. Cambridge, UK. constraints are independent of the players' propensities to guilt.