Discussion Papers

Promises and Conventions – A Theory of Pre-play Agreements

Topi Miettinen
University of Helsinki, RUESG, HECER
and University College London

Discussion Paper No. 97

ISSN 1795-0562

HECER – Helsinki Center of Economic Research, P.O. Box 17 (Arkadiankatu 7), FI-00014 University of Helsinki, FINLAND, Tel +358-9-191-28780, Fax +358-9-191-28781, E-mail info-hecer@helsinki.fi, Internet www.hecer.fi

HECER
Discussion Paper No. 97

Promises and Conventions – A Theory of Pre-play Agreements*

Abstract:

Experiments suggest that communication increases the contribution to public goods (Ledyard, 1995). There is also evidence that, when contemplating a lie, people trade off their private benefit from the lie with the harm it inflicts on others (Gneezy, 2005). We develop a theory of bilateral pre-play negotiation that assumes the latter and implies the former (alternatively we can interpret the agreement as a convention). A preference for not lying (transgressing conventions) provides a partial commitment device that enables informal agreements. We establish some general properties of the set of possible agreements and characterize the smallest and largest such set. In symmetric games, pre-play agreements crucially depend on whether actions are strategic complements or substitutes. With strategic substitutes, commitment power tends to decrease in efficiency whereas the opposite may be true with strategic complements. Also this finding is supported by experimental evidence.

JEL Classification: C72, C78, Z13.

Keywords: pre-play negotiation, communication, agreements, conventions, norms, guilt.

Topi Miettinen

Department of Economics
University of Helsinki
P.O. Box 17 (Arkadiankatu 7)
FI-00014 University of Helsinki
FINLAND

e-mail: topi.miettinen@helsinki.fi

*The author thanks *the Research Unit of Economic Structures and Growth (RUESG)* at the University of Helsinki and *the Yrjö Jahnsson Foundation* for financial support. Also, I am grateful to Steffen Huck for advice, encouragement, suggestions, and discussions and also to Philippe Jehiel on the same grounds. Finally, I would like to thank Martin Dufwenberg, Daniel Friedman, Antonio Guarino, Klaus Helkama, Klaus Kultti, Mikko Leppämäki, Liisa Myyry, Matthew Rabin, Francesco Squintani, Emma Tominey, Joel Sobel, Pekka Sääskilahti, Juuso Välimäki and the seminar participants at CEBR in Copenhagen, Game Theory Festival at Stony Brook, Helsinki Center of Economic Research, University College London, Zeuthen Workshop in Copenhagen. All errors are mine.

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

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

1 Introduction

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

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

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

Two findings in experimental economics give a reason to believe that the magic pots and the dislike to breach oral agreements are worth taking seriously: First, communication increases contributions in public good games[1]. Second, if people lie, they tend to dislike it; and the more harm they inflict on others by doing so, the more they seem to dislike it. This is shown by Gneezy (2005) and by several studies in social psychology. In public good games, agreeing to contribute more than one actually intends to contribute amounts to a lie which harms others. Thus, a theory that assumes the latter finding provides an explanation for the former finding.

Pre-play agreements by mutual consent provide a means to establish commonly known expectations about each others behavior. Having justified expectations of each others behvaior, players dislike letting down the opponent if the opponent does not. Aversion for letting down opponent's expectations is studied experimentally and theoretically by Dufwenberg and his co-authors in several papers. In this paper, we argue that for let-down aversion to emerge, the expectations must satisfy three criteria: they must be 1) coinciding, 2) commonly known and 3) justified. We argue that conventions and pre-play agreements are

[1]See Ledyard (1995) for a review of experimental research on public goods. This result holds for public good games without a threshold. The evidence that communication would increase contributions in public good games with thresholds is more mixed - the increase in contributions is not always significant.

natural devices that generate mutual expectations satisfying these criteria. We present a tractable model of let down aversion which takes into account these features.

In the model, the more players dislike breaching agreements the more harm they inflict on others by doing so. We show that in symmetric games with strategic substitutes such as the public good provision with a decreasing returns to scale technology, or the Cournot duopoly, there is a conflict between the efficiency of the agreement and the incentives to respect it. On the other hand, in an important class of symmetric games where actions are (weak) strategic complements (Bulow, Roberts and Klemperer, 1985), such conflict is circumvented: a symmetric efficient agreement can be made, if any. This class includes the public good provision game with constant returns to scale production technology and other team work and parnership designs as well as Bertrand duopolies with imperfect substitutes.

Public good experiments with communication lend strong support for our theory: Isaac and Walker (1988) adopt a constant returns to scale technology and find a strong positive effect of communication on efficiency. Average contribution levels are practically first-best efficient. Isaac, McCue and Plott (1985) adopt a decreasing returns to scale production technology. Despite the positive effect of communication on efficiency, they find that the average contribution levels are well below first-best efficient[2].

Our theory considers bilateral agreements in a wide array of strategic two-player interactions. The *underlying game,* the game that is played when an agreement is established, can be any normal form game. We assign the guilt cost properties that experimental and narrative research in economics and psychology has discovered. We assume that the general principles that govern guilt are the same for all players. Players may differ only in their *proneness to guilt*, i.e. how much weight they put on the guilt cost. We abstract from how an ageement is established (in pre-play negotiations, the negotiation protocol) but assume that the *agreement* is either an action profile of the underlying game or disagreement. Having an agreement on a profile, a player who breaches may feel guilty, which lowers her utility.

Given a game and players' proneness to guilt, each agreement maps the game into another game with the same strategy sets, but different payoffs. We are interested in which action profiles are *agreeable,* which action profiles can be enforced by guilt. Also, we are interested in how agreeability is affected by changes in (1) the underlying game, (2) the agreement, and (3) players' proneness to guilt.

Agreeability is defined in terms of incentive compatibility. An action profile is *incentive compatible* if neither player prefers breaching. That is, for any unilateral deviation from the profile, the guilt cost is larger than or equal to the underlying game benefit for the deviator. We call the difference between the

[2]These two studies are the only one's that allow subjects to play repeatedly and learn about the game. Actually, Isaac and Walker (1988) have one design with constant returns to scale technology and another with decreasing returns to scale technology. With the former design, first best efficiency is reached whereas the latter falls short of first best.

underlying game benefit and the guilt cost the *incentive to breach*.

Which agreements are agreeable will depend crucially on the properties of the guilt cost. We adopt the following properties, which are based on stylized facts in research in social psychology and experimental economics[3]:

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

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

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

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

Property {A} captures the idea that if my breaching the agreement causes my opponent to lose a toe, I do not suffer more than if my breaching the agreement causes my opponent to lose a leg. Gneezy (2005) finds strong support for property {A}: his experiments suggest that people trade off the benefits of lying against the harm that lying inflicts on the opponent. Property {B} is a no-sucker property: I will not feel guilty about breaching an agreement if my opponent breaches the agreement, too. According to property {C} the agreement's generosity induce stronger guilt. In pre-play negotiations, since there is guilt only if the opponent does not breach the agreement, the fact that the opponent respects and the fact that the agreed payoff is high indicate that the opponent is kind and generous. Hence, breaching the agreement and not reciprocating this will induce stronger guilt than if the agreement had been less generous. In the context of conventions, this effect may be weaker but even there a player may be more willing to breach a convention she considers unjustified. Properties {B} and {C} render guilt reciprocal. They also emphasize the idea that players are less willing to let down justified expectations. Property {D} expresses the idea that if the expectations do not coincide - there is no agreement, there is no guilt of breaching.

If the agreement is established by pre-play negotiations, it is natural to think that each player can veto any agreement. We say that an action profile is *individually rational* if it ensures that each player gets more than in her least preferred Nash equilibrium of the underlying game. In pre-play negotiations, upon deciding whether to signal disagreement, each player acts as if she knew that doing so will imply coordination on her worst Nash equilibrium.

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

Further towards our main conclusion, we find unambiguous effects on the incentive to breach when the terms of the agreement are altered (if the agreement

[3]In addition to their intuitive appeal, we present experimental evidence and psychological theory that supports these assumptions in section 2.

is agreeable in the first place): *in symmetric games with strategic complements, changing either agreed action so as to improve a player's agreed payoff decreases her marginal incentive to breach.* These effects are quite natural and intuitive: if the terms of the agreement are better for me, I have a lower incentive to breach. Yet, the result does not hold generally.

In *symmetric games with strategic substitutes*, as far as changes in player's own action are concerned, the player's payoff and her incentives to respect agreements are still naturally aligned. Yet, changing the opponent's agreed action implies quite the opposite effect: *the marginal benefit increases and the marginal harm on the opponent decreases when the opponent's action is changed so as to improve player's payoff.* This is the source of our result, identifying a conflict between efficiency and incentives in symmetric games with strategic substitutes, such as the standard Cournot duopoly or public good provision with a decreasing returns to scale production technology.

We also describe the agreeable set in a more general class of games and characterize the smallest and largest such set: Nash equilibria are always agreeable and nothing but Nash equilibria are agreeable for players with no proneness to guilt. Yet, a player who is sufficiently prone to guilt can agree on any individually rational profile that she cannot alone Pareto-improve and strictly benefit herself.

The paper is organized as follows. Section 2 presents related literature in economics and psychology. Section 3 presents the model. Section 4 studies a public good game. Section 5 presents general results and section 6 studies games with ordered strategy spaces. Section 7 considers a Cournot duopoly example. Section 8 concludes and discusses some further research problems.

2 Related literature

Economics. Evidence from experiments in public good games shows that even without communication subjects contribute positive amounts when purely monetary incentives make zero contribution a strictly dominant strategy. Existing social preference models nicely capture this effect. Yet, a largely unexplained finding is that communication raises the contributions well above the amounts observed without communication (Ledyard, 1995). Earliest experiments show this in prisoner's dilemma games (Loomis,1959; Radlow and Weidner, 1966). Recent studies for the two-person prisoner's dilemma case are provided by Duffy and Feltowich (2002) and (2005)[4].

A way forward in explaining the effect of communication is to combine one of the inequity aversion theories (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) with Farrell's (1987) idea that agreements will be stuck by if there is no incentive not to do so. Yet, this fusion of theories can only account for the experimental findings as long as the payoffs are not too asymmetric, since if they

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

are, symmetric contribution profiles lead to unequal payoffs and players with payoffs below the average cannot commit to these profiles. Even in symmetric environments, if the more efficient symmetric equilibria exist in the underlying game, the learning process never reaches these equilibria when communication is not present and, yet, these outcomes are reached when communication is allowed for (Isaac, McCue and Plott (1985); Isaac and Walker (1988)). Gneezy (2005) and Ellingsen and Johanneson (2004), on the other hand, carrie out further communication experiments and find behavioral patterns that cannot be explained by inequity aversion theories alone but which point to a preference for not lying.

The extensive form extension of Rabin's (1993) theory of reciprocity as introduced by Dufwenberg and Kirchsteiger (2004) is another candidate for capturing the phenomenon. Nevertheless, Charness and Dufwenberg (2003) show that sequential reciprocity cannot fully account for the detected behavioral patterns related to communication. They conclude that there must be a separate preference related to lying and introduce, independently of the contribution in this paper, the guilt-aversion equilibrium, where a player suffers a cost when she acts counter to the opponent's expectation on her behavior[5]. 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. Nevertheless, the role of communication is only implicit in their model. Furthermore, in their model, however unjustified the opponent's expectation is, guilt is constant whenever the harm on the opponent is the same.

Our model can be considered as an enriched and tractable CD model. Players dislike breaching an agreement only if she expects that the opponent does not act counter to it and only if the expected behavior in the agreement treats the player sufficiently well. Further, we are explicit about the effect of communication and the agreement. This view is supported by experimental evidence: Lev-on (2005) reviews communication experiments in public good games and concludes that mere identification or discussion which lacks explicit promising loses some of its effectiveness in supporting cooperation[6]. 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 negotiations and conventions.

Guilt has been discussed in several papers since Frank (1988) who argues

[5]Thus, like the theories of reciprocity, the theory falls into the category of psychological game theory (Geanokoplos, Pierce, and Stachetti 1989) where players' payoffs depend on beliefs explicitely. See also Dufwenberg (2002). Our model can be considered as a tractable model of let-down aversion where a player may be averse to act counter to a *justified* expectation. See Miettinen (2005).

[6]Furthermore, mere face-to-face identification increases cooperation especially in simple prisoner's dilemma games where coordination on group optimum is easy (Bohnet and Frey (1998)). Yet, cooperation rates are significantly weaker than when interactive communication is allowed for.

[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 (2002).

that it may well be materially profitable for an agent to have a conscience - a dislike for disobeying social norms. A recent model on emotional cost of breaching social norms is provided by Huck, Kubler, and Weibull (2003). These models involve no communication. Ellingsen and Johanneson (2004) do allow for communication and study the interplay of inequity aversion and guilt in a specific hold-up problem between a seller and a buyer. Their model is similar to ours in that guilt does not depend on the beliefs explicitly. Also, guilt is suffered if one breaches a promise. However, their model of guilt is simpler, since it does not take into account the reciprocal elements of opponent's behavior and it assumes that breaching a promise inflicts a constant guilt cost.

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

As far as property {B} is concerned, Baumeister, Stillwell, and Heatherton (1995) find that people feel more guilty about transgressions involving an "esteemed" person than about transgressions involving someone they hold in low regard. It is rather appealing to suppose that, if the opponent breaches the agreement, the esteem of a 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 opponents are likely to be esteemed. According to Clark and Mills (1984) and Clark (1979), concern for the other's welfare is the defining feature of communal relationships as opposed to exchange relationships. According to Baumeister, Stillwell, and Heatherton (1995), guilt is more likely to arise in the former type than in the latter type of relationships.

So as to property {D}, an agreement or an action-norm explicitly states an expectation and a standard of behavior for the play phase. Not reaching an agreement indicates players' inability to establish such a standard and a shared expectation. Millar and Tesser (1988) note that guilt depends on a concurrence of one's own expectations of behavior and those of the other person. Guilt appears mainly when there is a match in expectations of behavior. Such a match of expectations is established either by an exogenous action-norm or a pre-play agreement to an action profile. On the other hand, some 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 research 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 or a commonly known action-norm.

[8] See Ledyard (1995) and Pavitt and Shankar (2002).

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 reputation and repetition flavor 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 (1994), "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 (1994) 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 generalized to other relationships. ... In particular, a well-socialized individual would presumably have learned to feel guilty over inflicting harm to even a stranger.

Based on this view, we elaborate on the idea of guilt as an internalized punishment payoff in a repeated game prior to which players agree on a stationary pattern of play in the appendix.

In the present model, as in theories of fairness, players internalize 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.

3 The model

Let Γ be a two-player simultaneous move normal form game, below referred to as the *underlying game*.[9] Before the game is played, an agreement - a mutual expectation - is establihed either by pre-play negotiations or by convention. Generally, the pre-play negotiations may have an arbitrary strategic structure or the agreement may be exogenous - the only requirement is that there is an agreement or disagreement on how to play [10].

We rule out the use of mixed strategies in the underlying game. If we allowed for mixed strategies, we should determine whether guilt is a function of consequences only or whether guilt is felt even if a mixed strategy different from the agreed one is chosen but the random draw picks up a pure strategy that is in the support of the agreed mixed strategy[11].

[9] The theory allows for a straightforward extension to sequential two stage games.

[10] *Preplay negotiation* is a finite extensive form game tree. The terminal histories are associated with an oral (non-binding) agreement, or with disagreement.

[11] On the other hand, we could easily allow for correlated strategies where players agree on

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 *action profile* $s = (s_i, s_j) \in S = S_i \times S_j$. The *underlying game payoff* of player i is $u_i(s)$. Notice that this payoff may well include social preference terms.

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

The negotiations or the convention establishes an agreement, m, on how to play, or disagreement. Thus, we restrict $m \in S \cup \{d\}$ where d denotes 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*[12] indicates how much more than u_i^* the player gets if both respect the agreement, $v_i(m) \doteq u_i(m) - u_i^*$. If player i deviates from the agreement, we get the *harm* on j by subtracting j's payoff at the deviation profile from the payoff at the agreed action profile, $h_j(m, s_i) \doteq u_j(m) - u_j(m_j, s_i)$. Similarly, i's *benefit from breaching* is $b_i(m, s_i) \doteq u_i(m_j, s_i) - u_i(m)$.

In this paper, we restrict focus to simultaneous move games. Notice, that we could easily extend our theory to corresponding Stackelberg games, say, with player one the leader and player two the follower. That player one moves first gives her perfect commitment power. If the leader breaches, the follower does not suffer from guilt and her payoff coincides with the UG payoff. Thus, the follower will choose an UG best reply to the leader's action. In the Stackelberg version of the theory, we should replace the worst Nash payoff with the worst Stackelberg payoff.

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*, $g_i(v_i(m), h_j(m, s_i))$, depends on the inflicted harm and on the agreed payoff. The utility function over the action profiles in the entire game is assumed to be additively separable in guilt and the underlying game payoff.

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

The entire game payoff now depends on m and, due to guilt, talk is not cheap. The guilt cost is represented by $\theta_i g(v_i(m), h_j(m, s_i))$ which is assumed to be

a given random draw on how to play: guilt would be a function of the expected agreed payoff.

[12] Most of our results would be unaltered if we alternatively suppose that the reference point in the agreed payoff is the player's worst Pareto-efficient Nash payoff which is the lower bound for a long pre-play negotiation payoff derived in Rabin (1994).

non-negative. This rules out revengeful feelings or spite, on the one hand and positive emotions related to respecting agreements, on the other hand. This is somewhat restrictive, but here we want to focus on guilt.

The parameters $\theta = (\theta_1, \theta_2)$ capture players' *proneness to guilt*. For a given deviation, a player with a higher proneness to guilt suffers more. We only allow for non-negative proneness to guilt, $\theta_i \in [0, \infty)$. If it is common knowledge that the proneness to guilt of both players equals zero, the model coupled with a communication protocol is one of cheap talk[13].

Notice first, that the guilt cost depends on the agreement and on the deviation only indirectly through the agreed payoff and the harm. Second, choosing the agreed action m_i minimizes the guilt cost at the second stage. Furthermore, (BD) implies that if disagreement is reached, then there is no guilt cost. We assume that each player can unilaterally enforce disagreement, d. Also, there are no bad feelings about own cheating if the opponent cheats too. Thus (BD) introduces properties {B} and {D} into the guilt cost.

Moreover, we assume, that the guilt cost $g(v_i, h_j)$ is weakly increasing in the agreed payoff and in the harm. This introduces properties {A} and {C} into the guilt cost.

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

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

Also, we assume that if the player causes no harm to the opponent[14] 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 positive, then the guilt cost is strictly positive:

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

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

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

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

[13] As in Farrell (1987) but with any finite extensive form communication protocol ending up in an agreement - an action profile of the game.

[14] 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 (or its straightforward extension to sequential two-stage games) predicts rejection conditional on refund request and return rate above one which the data in Andreoni seems to confirm.

is [15]

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

This function is zero if the harm is non-positive or if the agreed payoff is below 0. Otherwise, it is strictly positive. It is increasing in the harm and in the agreed payoff.

We suppose that the proneness to guilt types and the language are common knowledge. Thus, players have correct point predictions about their opponent's proneness to guilt and beliefs of all degrees coincide. Also, players do not have to worry that the opponent might interpret an agreement to 'meet at noon' as an agreement to 'meet at quarter past noon.' Both these considerations are relevant but at this first step we abstract from this.[16]

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 and players' proneness to guilt is given by θ. Denote by $s^*(m;\theta) = (s_i^*(m;\theta), s_j^*(m;\theta))$ the equilibrium correspondences in that subgame.

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

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

and

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

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 the harm inflicted on the other when i breaches but not j. Player's *incentive to breach* an agreement m is the difference between the benefit from breaching and the guilt cost, $B_i(m, s_i; \theta_i) \equiv b_i(m, s_i) - \theta_i g(u_i(m), h_j(m, s_i))$.

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 holds 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 no player prefers enforcing disagreement (pre-play negotiations) over playing m, i.e. if for $i = 1, 2$

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

[15] The entire game preferences of this form with $\gamma = \varphi = 1$ belong to the class of Cox-Friedman (2002) preferences with $\alpha = 1$ with the emotional state depending on the agreed payoff $v_i(m)$.

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

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

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

4 A public good game

The prisoner's dilemma is a stylized version of a public good game where it is strictly dominant not to contribute:

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

where $h_i > u_i > 0$ and $b_i > 0$ for $i = 1, 2$. Supposing that the guilt cost takes the simple form of the example given in (2), player i respects an agreement to contribute, $m = (C, C)$, (given that the opponent does) if and only if

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

This is an incentive compatibility condition. Moreover, both contributing is individually rational by the structure of the prisoner's dilemma. So, an agreement on (C, C) should be particularly easy to reach if b_i is small and h_j is large - just as Gneezy (2005) 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, N) and (N, C). Both not contributing is incentive compatible and individually rational for all types since it is the unique Nash equilibrium. Hence, (N, N) is always agreeable and (C, C) is agreeable if (5) holds for both players.

Proneness to guilt may transform a prisoner's dilemma into a coordination game. This is a familiar property of fairness models. Yet here, first, the transformation is explicit; and second, the ability to commit to contribute 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 contribute than when they do not, $u_i - 0$. On the other hand, the payoff of the opponent is internalized only to the extent of how much a player's defection affects the opponent's payoff.

Guilt inflicts a cost of defection. It is trivial that if this cost is sufficiently large to balance off the benefit from breaching, the player can credibly commit not to defect. Yet, the prisoner's dilemma is a rather degenerate game: there is only one action profile that Pareto dominates the underlying game Nash equilibrium. Thus, the set of agreements under negotiation is very limited.

Our pre-play negotiations model may have bite in any game with an inefficient equilibrium.

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

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

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

Let us suppose for the time being that the guilt cost is given by (2) with $\gamma = \varphi = 1$ and let the production technology have constant or decreasing returns to scale, $G'' \leq 0$. Players can agree to any agreement where both get a positive payoff and the guilt is sufficient to prevent breaching. The harm due to a unit underprovision reads $h_j(m, m_i - 1) = G(\sum m_i) - G(\sum m_i - 1)$ which is decreasing in the sum of contributions and thus in efficiency when too little is contributed. The marginal benefit from a unit underprovision vis-à-vis the agreement is $1 - h_j(m, m_i - 1)$ and thus it is increasing in the sum of contributions.

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

$$\begin{aligned} &1 - G(\sum m_i) + G(\sum m_i - 1) \\ &-\theta_i[G(\sum m_i) - G(\sum m_i - 1)]^{\varphi} \max\{u_i(m), 0\}^{\gamma}. \end{aligned} \tag{6}$$

Supposing that an indifferent player respects the agreement, a player will breach if and only if (6) is positive. This is the marginal incentive compatibility condition. Since benefit from breaching is increasing and the harm on others is decreasing in the sum of contributions, there is a conflict between the efficiency of the agreement and the incentives to respect.

The agreement must also satisfy individual rationality,

$$u_i(m) = G(\sum_{k=1,2} m_k) + 10 - m_i \geq 0. \tag{7}$$

Notice yet, that (7) is actually redundant: it is implied by the incentive compatibility condition. When the agreed payoff approaches zero, the second term

of (6) approaches zero too and an agreement off the underlying best reply correspondence is not incentive compatible for any type. Thus the marginal incentive compatibility conditions are necessary and sufficient for agreeability (6).

Some of the properties explicit in (6) are worth emphasizing. To a lesser extent, a player with a higher proneness to guilt can agree on a larger set of agreements. Second, the relative contributions matter (but not the relative payoffs). More importantly, as identified above, the trading off of marginal harm and marginal benefit implies a conflict between efficiency and incentives when $G'' < 0$. If we isolate the effect of the own action, m_i, on the incentive to breach, the conflict is amplified by its positive effect on the agreed payoff. Yet, increasing m_j has the opposite effect on the agreed payoff and, since by assumption efficiency is increased, the overall agreed payoff effect on guilt is positive. This tends to decrease incentives to breach.

Thus, whether or not there is a conflict overall depends on G'' on the one hand and on G' and $\frac{\partial g}{u_i}$ on the other. If G'' is close to zero the trading off of benefit and harm is unaffected but the agreed payoff effect decreases incentives to breach. Yet if G'' is larger and the agreed payoff does not much affect guilt, the effect of trading off benefit and harm increases incentives to breach. Furthermore, if G'' is large the agreed payoff effect tends to fade away with efficiency. Eventually, if we have an interior group optimum, there will be a conflict between efficiency and incentives as we are sufficiently close to the group optimum.

Yet, as a special case, if there are constant returns to scale, $G' = \alpha$, the marginal payoffs are constant and the changes in breaching incentives are driven only by the agreed payoff effects: incentives to breach decrease with efficiency. Further, the marginal benefit from breaching decreases in α and the marginal harm increases in α and the agreed payoff of any agreeable action profile increases in α. Thus, it is easier for the players to agree when the marginal per capita return is higher.

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

Proposition 1 *Let g satisfy (2) with $\varphi \geq 1$. In the public good game,*

- *an agreement is agreeable iff the marginal incentive to breach is non-positive for $i = 1, 2$.*
- *player i's marginal incentive to breach is increasing in m_i.*
- *if $G' = \alpha$, player i's marginal incentive to breach is decreasing in α and in m_j and in $\sum_{k=1,2} m_k$.*
- *if $G'' < 0$ and $\gamma = 0$, player i's marginal incentive to breach is increasing in m_j and in $\sum_{k=1,2} m_k$.*

Proof. For the first claim, it is straightforward that

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

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

$$
\begin{aligned}
& 1 - G(\sum_{k=1,2} m_k) + G(m_j + m_i - s_i) - \theta_i g(v_i(m), [G(\sum_{k=1,2} m_k) - G(m_j + m_i - s_i)]) \\
\leq\ & [1 - G(\sum_{k=1,2} m_k) + G(m_j + m_i - 1)][m_i - s_i] && (8) \\
& -\theta_i g(v_i(m), [G(\sum_{k=1,2} m_k) - G(m_j + m_i - 1_i)](m_i - s_i)) && (9) \\
\leq\ & [1 - G(\sum_{k=1,2} m_k) + G(m_j + m_i - 1)] && (10) \\
& -\theta_i g(v_i(m), [G(\sum_{k=1,2} m_k) - G(m_j + m_i - 1_i)]) && (11) \\
\leq\ & 0 && (12)
\end{aligned}
$$

where the first inequality follows from the fact that the 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 $[m_i - s_i] \geq 1$.

For the second claim, increasing m_i will decrease $u_i(m)$ and thus $v_i(m)$, increase $b_i(m, m_i - 1) = 1 - G(\sum_{k=1,2} m_k) + G(m_j + m_i - 1)$ and decrease $h_j(m, m_i - 1) = G(\sum_{k=1,2} m_k) - G(m_j + m_i - 1)$.

For the third claim, increasing α will increase $u_i(m)$ and thus $v_i(m)$, decrease $b_i(m, m_i - 1)$ and increase $h_j(m, m_i - 1)$. Increasing m_j or $\sum_{k=1,2} m_k$ will increase $u_i(m)$ and thus $v_i(m)$ and leave $b_i(m, m_i - 1)$ and $h_j(m, m_i - 1)$ unaffected.

For the fourth claim, increasing m_j or $\sum_{k=1,2} m_k$ will increase $u_i(m)$ and thus $v_i(m)$ but since $\gamma = 0$ this will not affect g. Increasing m_j or $\sum_{k=1,2} m_k$ will increase $b_i(m, m_i - 1)$ and decrease $h_j(m, m_i - 1)$. ■

Proposition 1 establishes that instead of checking for all possible deviations it is necessary and sufficient simply to check for a local deviation. Thus, to determine a player's potential to agree, we can look for agreements where the player is indifferent between respecting and deviating marginally. Any agreement where a player's action is smaller or an opponent's action is larger than at the boundary is agreeable for that player.

Figure 1 shows the agreeable set for $G' = \alpha = \frac{3}{4}$, $\theta_i = 4$ and $g(v_i, h_j)$ as in (2).

The action profiles that belong to player one's potential to agree are marked with plus signs and the action profiles that belong to player two's potential to agree are marked with crosses. Thus the action profiles marked with asterisks are agreeable action profiles, $A(\Gamma_{PG}(\frac{3}{4}), (4, 4))$. Notice, that the best reply curves lie on the axes and that each player's best reply curve is agreeable for each player. Thus, the Nash equilibrium, $(0, 0)$, is agreeable. Notice also that some efficient action profiles are agreeable: for instance the symmetric efficient action profile where both give a full contribution, $m = (10, 10)$.

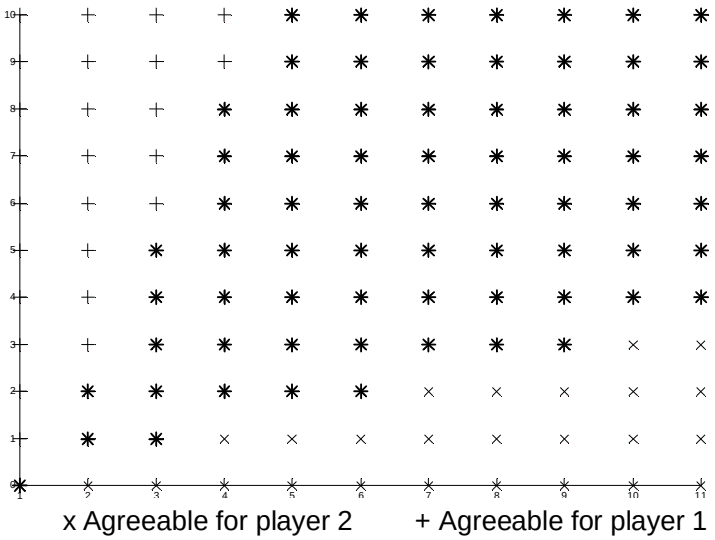

Figure 1: Figure 1: The agreeable set

Figure 2 illustrates how easy it is to agree on the symmetric efficient action profile. Specifically, it plots the critical θ that makes a player indifferent between breaching and respecting as a function of α. As stated above, increasing α makes the incentive compatibility constraint less stringent and, thus, the function is decreasing.

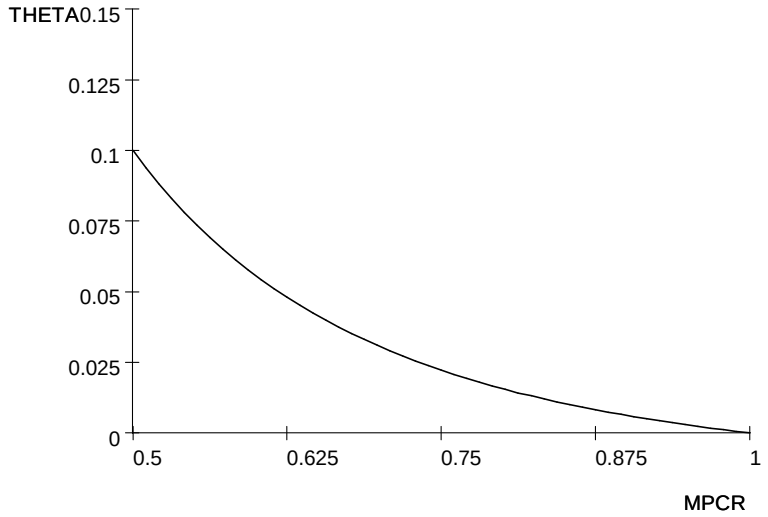

Figure 2: Indifferent player at symmetric efficient profile given α

Indeed, we have shown in this section that when communication is allowed in public good games and players are prone to guilt, players may agree to contribute positive amounts and guilt provides the necessary incentives to commit to the agreement. Further in regards to the experiments by Isaac, Mccue and Plott (1985) and Isaac and Walker (1988), we have suggested that a likely explanation for the differences in their results may not be that it is difficult for the players to identify an interior group optimum. Rather that agreements sufficiently close to the group optimum suffer from a conflict between efficiency of the agreement and the incentives to respect it. This conflict is absent when it is optimal to contribute everything to the public good as in Isaac and Walker (1988). Notice that it is crucial here that guilt cost is convex in the harm on the other. For instance constant guilt cost due to a deviation, (1), cannot account for the difference since with that specification guilt is concave in the harm on other[17].

Other social preference models can explain positive contributions to public goods but none have explained why communication further increases contributions. In this section we have shown this for players who without communication have a strictly dominant strategy to contribute nothing. This does not seem to comply with the empirical finding that even without communication positive amounts are contributed. Yet, the next section develops the theory in the more general case where the underlying game preferences may take an arbitrary form (and may thus involve social preferences) and equilibria of the game are inefficient. The section shows in particular how the present theory can account for

[17]This will imply, the model of Ellingsen and Johanneson (2004) cannot account for the differences in efficiciency results of Isaac, McCue and Plott (1985) and Isaac and Walker (1988).

the fact that communication increases contributions from the levels that prevail without communication.

Further, the next section generalizes the sharp contrast in feasibility of first–best efficiency between constant returns to scale technology and decreasing returns to scale technology in public good games: there is a conflict between incentives and efficiency in symmetric submodular games where payoff is monotone in opponent's action. Such a conflict tends to be absent in symmetric supermodular games with monotone payoff in opponent's action.

5 Properties of the agreeable set

This section derives some simple properties that apply to any normal form underlying game. First, any UG Nash equilibrium is agreeable. Thus, the agreeable set is never smaller than the set of Nash equilibria of the UG. Second, a Nash equilibrium remains a Nash equilibrium of most subgames that follow an agreement. Yet, if an agreement is such that a player can unilaterally deviate to an UG Nash equilibrium, then this UG Nash equilibrium may no longer be a Nash equilibrium when the agreement is made. Third, if a player can deviate from an agreement and thereby benefit both players, the action profile is not agreeable. Yet, any individually rational profile that does not satisfy this property can be agreed upon if proneness to guilt is sufficiently high. This characterizes the largest possible agreeable set as opposed to the smallest such set - the set of UG Nash equilibria.

In the prisoner's dilemma, both (D, C) and (D, D) are agreeable for the row player. Both profiles are individually rational for the row player and the row player's agreed action is an UG best reply to the agreed action of the column player. Underlying game preferences drive the player to choose the best reply. If a player's agreed action is a best reply to the agreed action of the other player, the guilt cost of deviating would only add to the forgone UG payoff . The first part of the following lemma establishes this general finding.

On the other hand, if a player's agreed action is not a best reply to the opponent's agreed action, then the agreement belongs to the player's potential to agree if and only if it is incentive compatible. The UG benefit from breaching is positive at least for the deviation to the best reply; if individual rationality is violated, the guilt cost is zero and the agreement is not incentive compatible.

Lemma 2 *Let $m_i \in BR_i(m_j)$. Then $m \in A_i(\Gamma, \theta_i)$ iff (IR_i) holds.*
Let $m_i \notin BR_i(m_j)$. Then $m \in A_i(\Gamma, \theta_i)$ iff (IC_i) holds.

Proof. See appendix. ■

This lemma is useful for characterizing each player's potential to agree: on the best reply curve, all individually rational agreements are agreeable. Off the best reply curve, all incentive compatible agreements are agreeable and no other agreement is. Thus, for non-equilbirium conventions only incentive compatibility matters. On the other hand, lemma 2 enables us to generalize the finding that, in the prisoner's dilemma, the defection equilibrium is agreeable for any

proneness to guilt types. By definition, any Nash equilibrium payoff is individually rational. Thus by the first part of lemma 2, any Nash equilibrium belongs to each player's potential to agree. Thus, a Nash equilibrium is agreeable.

Proposition 3 *If $m \in NE(\Gamma)$, then $m \in A(\Gamma, \theta)$.*

Proof. See appendix. ■

First, for zero proneness to guilt types, Nash equilibria are the only agreeable action profiles.[18] Second, 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 profiles that are agreeable.

Recall that we ruled out mixed strategies and thus an agreeable profile may not exist. Notice, that allowing for mixed strategies would ensure that an agreeable profile always exists (whichever way we think about guilt): an underlying game Nash equilibrium is always agreeable and with mixed strategies a Nash equilibrium always exists in finite games.

Yet, pre-play negotiations may create an equilibrium selection problem when there is an agreement in place and players are prone to guilt. For instance, when players agree on cooperation in the prisoner's dilemma, defection remains an equilibrium of the transformed game. 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 $\Gamma(m;\theta)$. This shows that even if m is a Nash equilibrium of $\Gamma(m;\theta)$, there may be other equilibria as well.

Lemma 4 *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. ■

The equilibrium selection problem apparent in lemma (4) is avoided however if we suppose that players will conform to the agreement, if there is no incentive not to do so, as assumed in Farrell (1987).[19] Lemma (4) shows that an UG Nash equilibrium may be a Nash equilibrium of a subgame where players do not agree on that Nash equilibrium. Notice yet, that this is not true for any agreement. Nash equilibria may be removed from the game.

Consider the following game of chicken:

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

(13)

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

[19] Applying Farrell (1987), we may refine 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, $s^*(m;\theta) = m$.

Farrell and Rabin (1996) 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.

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

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

Lemma 5 *For any m, if there is a player i such that there exists s_i such that $u_i(s_i, m_j) > u_i(m)$ and $u_j(s_i, m_j) \geq u_j(m)$ then $m \notin A(\Gamma, \theta)$ for any θ.*

Proof. See appendix. ■

Lemma 5 follows immediately from the monotonicity (AC) and the strict cost (EF) conditions: when the harm inflicted on the other is non-positive, there is no guilt cost. Since a player can make herself better off, she will do so and the agreement is not incentive compatible.

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

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

(14)

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 than in (B, L). 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 (1987). Indeed, this type of plurality may question whether (B, L) is agreeable in the first place. But for our analysis, it is sufficient to notice that since player 1 can make both better off, the agreement is not incentive compatible.

In (14), players cannot agree on (T, R) either, since player 1 gets a smaller payoff than in the underlying game equilibrium, (B, R). On the other hand, if player 2's proneness to guilt is small, players 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 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, if UG payoffs are finite, with sufficiently high proneness to guilt all individually rational profiles are agreeable for which a Pareto-improving deviation does not exist (the deviator must strictly benefit), and no other profile is.

Proposition 6 *Let the underlying game payoffs be finite. Let* $v_i(m) > 0$ *for* $i = 1, 2$. *Then* $m \in lim_{\theta_1 \to \infty, \theta_2 \to \infty} A(\Gamma, \theta)$ *iff for* $i = 1, 2$ *and for all* s_i, $u_i(m) \geq u_i(s_i, m_j)$ *or* $u_j(m_j, s_i) < u_i(m)$

Proof. See appendix. ■

If the set of Nash equilibria is the smallest set that is agreeable (cheap-talk), proposition 6 describes the largest possible agreeable set, the agreeable set for types that are infinitely prone to guilt.

Lemma 5 has another implication, which is mentioned here without a proof. Namely, within the agreeable set, the interests of the players are opposed for any change in one of the agreed actions.

Corollary 7 *Let* $(m_i, m_j), (m_i', m_j) \in A(\Gamma, \theta)$ *then*

$$\begin{array}{l} u_i(m_i, m_j) > u_i(m_i', m_j) \Rightarrow u_j(m_i', m_j) > u_j(m_i, m_j) \\ u_j(m_i', m_j) > u_j(m_i, m_j) \Rightarrow u_i(m_i, m_j) \geq u_i(m_i', m_j) \end{array} \tag{15}$$

6 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 generalize two results gained there: First, that the non-positive marginal incentives to breach are necessary and sufficient for a strategy profile to be agreeable. Second, trading off the marginal harm of a deviation with its marginal benefit implies a conflict between efficiency and incentives to respect when there are decreasing returns to scale in the public good production whereas such a conflict is absent with constant returns to scale. We show that, when the guilt cost is convex in the harm, the first result generalizes to underlying games with concave payoff functions in each action. For the second result, there is a conflict between incentives and efficiency in symmetric submodular games where payoffs are monotone. Such a conflict tends to be absent in symmetric supermodular games with monotone payoffs in opponent's action.

Symmetric games with strategic complements avoid the conflict only if certain convexity properties are present: a symmetric efficient agreement can be made if the players can agree on any symmetric profile that is not an underlying game equilibrium.

We now adopt some new concepts and notational simplifications. We denote the action s_i^n by its order label n so that for $k \in \mathbb{Z}$, $s_i^n + k \doteq s_i^{n+k}$. Also for $s \in S$ we let $s + k \doteq (s_i + k, s_j + k)$. We let the marginal benefit from breaching be defined as $\beta_i(m_i, m_j) \doteq b_i(m_i, m_j, m_i - 1)$, and the marginal harm as $\eta(m_i, m_j) \doteq h_i(m_i, m_j, m_i - 1)$. Thus $\beta_i(m+k) = \beta_i(m_i + k, m_j + k)$, $\eta_i(m + k) = \eta_i(m_i + k, m_j + k)$, and $u_i(m + k) = u_i(m_i + k, m_j + k)$ for $k \in \mathbb{Z}$.

We first set the scene by making *further assumptions on the underlying game.* 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. That is, for all s

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

and for all s

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

{3} The payoff functions are supermodular (so that actions are strategic complements). That is for all s

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

These properties are satisfied in the public good game, but in a degenerate manner: for all s, $\delta_i(s) = \sigma_i(s) = \phi_i(s) = 0$. The first assumption is not crucial. Indeed, if we reverse the ordering of strategies of both players, the payoff will be decreasing in opponent's action and, yet, concavity and supermodularity of the payoffs are unaltered. Thus, symmetric games with decreasing payoffs in opponent's action can be analyzed using the same artillery.

Also, we make *further assumptions on the guilt cost.* We assume that if it is convex in the harm, h_j, and in the agreed payoff, v_i, and that it is supermodular in its two arguments

{4} g is convex in h_j

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

Notice that the fact that the payoff is concave in the opponent's action implies that the harm h_j is a convex function of s_i, since the harm is just a rescaled negative of the underlying game payoff. Thus, by assumption {4}, the guilt cost is convex in s_i as a composite of two convex functions. Notice that assumption {4} rules out constant guilt cost, (1), for instance, since with that specification guilt is concave in harm[20]. 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 optimization problem. Hence, checking that neither prefers to breach the agreement marginally is necessary and sufficient for an agreement to be incentive compatible.

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

Definition 8 *(Marginal incentive to breach)*
If $u_i(m_i-1, m_j) - u_i(m) \geq 0$

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

[20]This will imply, the model of Ellingsen and Johanneson (2004) cannot account for the differences in efficiciency results of Isaac, McCue and Plott (1985) and Isaac and Walker (1988).

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

$$\mathbb{B}_i(m, \theta_i) \doteq \beta_i(m)$$

When $u_i(m_i - 1, m_j) - u_i(m) < 0$, there is certainly no incentive to deviate downwards but there may be an incentive to deviate upwards. The fact that $\mathbb{B}_i(m, \theta_i)$ does not involve any guilt cost when $u_i(m_i - 1, m_j) - u_i(m) < 0$ is due to the fact that, by assumption {1}, an upward deviation does not make the opponent worse off and thus the player does not suffer from guilt.

Consequently, assumption {1} on the underlying game payoffs together with lemma 5 gives us a necessary condition for an action profile to be agreeable. The play must belong to the following set[21]

$$M_F = \{m | u_i(m_i, m_j) \text{ is non-increasing in } m_i \text{ for } i = 1, 2\} \tag{16}$$

Next, we establish the necessary and a sufficient condition for agreeability that generalizes our finding in the public good game. We established above that, due to the convexity of the problem, there is no incentive to breach the agreement at the margin iff and only if there is no incentive to breach at all. Second, incentive compatibility implies individual rationality when off the underlying game best reply curves by lemma 2. Thus, we have the following.

Proposition 9 *Let* Γ *be finite. Let* $m_i \neq BR_i(m_j)$ *and* $m_i \notin \{s_i^1, s_i^n\}$. *Let {1}, {2} and {4} hold. Then an action profile is agreeable for* i *if and only if the marginal incentive to breach is non-positive.*

Proof. See appendix. ■

As the terms of the agreement are altered, the marginal incentive to breach is affected through three channels: i) the direct effect through the marginal benefit from breaching; ii) an indirect effect through the marginal harm on the opponent; iii) an indirect effect through the agreed payoff. The latter two are indirect in that they affect the marginal incentive to breach through the marginal guilt cost.

In the public good game, we found that the marginal incentive to breach is monotone in each agreed action. We can generalize this property. Let us first consider how a change in one agreed action affects the trading off of benefit and harm from breaching.

Let us start with the effect of the agreed action of player i, m_i. It is necessary that an agreeable action profile lies in M_F. But within M_F, the player's payoff must be decreasing in her action. Thus, the effect of a player's agreed action on her marginal benefit from breaching is nothing but the negative of the second derivative, $-\delta$. Thereby, increasing a player's agreed action increases her marginal benefit from breaching. Similarly, the effect of m_i on η_j is simply the second derivative, σ, since the harm is itself a rescaled negative of u_j and breaching takes place downwards. Thus increasing m_i increases β_i and decreases

[21] Except for $m_i = s_i^n$ of course.

η_j and both these effects have a positive impact on the marginal incentive to breach.

The effect of m_j on β_i and η_j rests on the strategic complementarity[22] of actions. Due to strategic complementarity, if the opponent increases her action, then the player has a stronger incentive to increase her own action. Since breaching takes place downwards, increasing the opponent's agreed action dampens the underlying game benefit from breaching. On the other hand, the higher the opponent's action, the more harm is inflicted on her by marginally decreasing the own action. Strictly supermodular games, where $\phi > 0$, constitute a set of games where such complementarities are present. The following lemma summarizes the effects of changing the terms of the agreement on the trade-off between the benefit and the harm of breaching.

Lemma 10 $\beta_i(m_i + 1, m_j) - \beta_i(m_i, m_j) = -\delta_i(s)$
$\eta_j(m_i + 1, m_j) - \eta_j(m_i, m_j) = \sigma_j(s)$
$\beta_i(m_i, m_j + 1) - \beta_i(m_i, m_j) = -\phi_i(s_i, s_j + 1)$
$\eta_j(m_i, m_j + 1) - \eta_j(m_i, m_j) = \phi_j(s_j + 1, s_i)$

Proof. See appendix. ■

Now consider the third effect - the agreed payoff effect - of m_i and m_j on the marginal incentive to breach. This effect goes through the agreed payoff. Corollary 7 together with {1} imply that the agreed payoffs change monotonically in the agreeable set: increasing own agreed action decreases the agreed payoff and increasing the opponent's action increases payoff. Thus, when m_i is increased, also the agreed payoff effect has a positive impact on the marginal incentive to breach. On the other hand, there is a negative impact when m_j is increased. Thus the agreed payoff effects are aligned with the marginal harm and benefit effects. Thus, in supermodular games, increasing an opponent's action decreases the marginal incentive to breach [23]. Similarly, increasing the own agreed action increases the marginal incentive to breach.

Proposition 11 *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 set.*

Proof. See appendix ■

Notice that the agreed payoff reflects a player's preference ordering of agreements conditional on both respecting. Thus, keeping one of the actions fixed and changing the other, the preference over agreements and the incentive to respect them are aligned. Yet, in symmetric submodular games where $\phi < 0$, there is some conflict in the preference over agreements and the incentive to respect them.

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

[23] Also, 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.

Furthermore, this implies that, apart from the agreed payoff effect, in symmetric submodular games where {2} holds, efficiency and incentives to respect are in conflict. To see this, notice that no agreement where a player is required to choose an action smaller than her underlying game best reply is efficient. Symmetric profiles that Pareto-dominate the equilibrium are such that both agreed actions are higher than in equilibrium. But, increasing both actions by one step, increases the marginal benefit from breaching and decreases the marginal harm on the opponent. Thus, abstracting from the agreed payoff effect and only focusing on the trading-off of harm and benefit, the incentive to breach is increased.

Theorem 12 *Let* Γ *satisfy {1}, {2} and* $\phi < 0$. *Let* s^* *be its unique symmetric equilibrium with* $\beta_i(s_i^*, s_j^*) = 0$. *If* $u_i(s^* + k) - u(s^*) > 0$ *for* $k \in \mathbb{Z}$ *then* $\beta_i(s^* + k) > 0$ *and* $\eta_i(s^* + k) < \eta_i(s^*)$.

Proof. See appendix. ■

While theorem 12 establishes a conflict between efficiency and agreeability, we know on the other hand that in the public good game efficiency and the incentives to breach may well be aligned: an efficient action profile can be agreed upon if and only if an interior non-equilibrium action profile can be agreed upon. In theorem 13, we establish that this holds more generally in symmetric supermodular games. The result is not as robust as the conflict result, however. We need some further, not very restrictive assumptions which are satisfied in many examples.

Either we need to suppose that guilt is unaffected by the agreed payoff ($\gamma = 0$ in the public good example above) or we suppose that the UG payoff is convex in identical changes of both actions. For the latter case, when the UG payoff is convex in this way and the payoff is increasing in such changes, it is increasing in symmetric changes from the symmetric interior equilibrium up to the symmetric efficient profile where actions cannot be increased any further. Thus, a symmetric non-equilibrium action profile is agreeable if and only if an efficient action profile is.

This argument suffices for the case that best reply curves are not particularly steep. When they are steep, there may be multiple equilbria and we can use Milgrom and Roberts (1990) result that in supermodular games when payoffs are increasing in opponent's action, the equilibria are ordered in terms of efficiency. Thus, the profile of maximal contributions is efficient and also agreeable as an underlying game equilibrium.

Theorem 13 *Let {1},{2} and {3} hold. Let* $\delta_i(s), \sigma_i(s)$ *and* $\phi_i(s)$ *be constant for* $i = 1, 2$. *Let* Γ *be symmetric and let* s^* *be its inefficient Nash equilibrium such that* $\beta_i(s^*) = 0$ *for* $i = 1, 2$. *Let* g *satisfy {4}. Suppose either (a) that* $\phi + \sigma \geq 0$ *and* $g(v', \eta) = g(v, \eta)$ *for all* η *and* $v', v > 0$ *or (b) that* $2\phi + \delta + \sigma \geq 0$ *and* g *satisfies {5}.*

Then, a symmetric efficient s *is agreeable iff a symmetric* $s \neq s^*$ *is agreeable*

Proof. See appendix. ■

Proposition 11 shows that in games with strategic complements the marginal incentive to breach has intuitive monotonicity properties: as the action of the opponent is increased, a player's incentive to breach decreases whereas the opposite is true when the player's own action is increased. On the other hand, in supermodular games, players are able to reach symmetric efficient agreements if anything else that is symmetric and tha is not an interior UG equilibrium can be agreed upon.

Notice again, that assumption {1} was made without loss of generality. All we need is symmetry. If the payoff is decreasing in the opponent's action, we can restore assumption {1} by reversing the ordering of each strategy set. This will affect neither the concavity of the UG payoff in each action nor the super- or submodularity of the underlying game payoff.

In addition to the linear public good game studied above, examples of symmetric supermodular games include, for instance, team work designs and partnerships, or the Bertrand duopoly with imperfect substitutes. Yet, as we have seen the monotonicity properties and efficiency results do not generally hold in symmetric submodular games where the payoff is increasing in opponent's action.

Examples of symmetric games with strategic substitutes are the game of chicken (see section 5) and public good provision with a concave production technology. The chicken is a stylized version of a public good game with a provision threshold. Experimental evidence on the effect of communication in the public good games with a threshold is mixed. On the other hand, Isaac, McCue and Plott (1985) find rather weak effects of communication on efficiency in a public good game with decreasing returns technology whereas Isaac and Walker (1988) find a very strong positive effect of communication on efficiency with a constant returns to scale technology. Thus, our theory organizes rather well the differences in the effects of communication in public good games.

The next section studies a Cournot duopoly as an example of a symmetric game with strategic substitutes. Thus, the incentives to respect more collusive agreements tend to be weaker.

7 Cournot duopoly

Let us now study an example to see what happens when supermodularity of the underlying game is violated. We transform a linear Cournot duopoly with imperfect substitutes where profits read as $\pi_i(q) = (\frac{19}{2} - \frac{1}{2}q_i - q_j)q_i$ and the strategy set is $q_i \in \{0, ..., 10\}$ into an equivalent game[24] where the strategy sets

[24]Notice that despite the negative strategies, this is indeed a game equivalent to a Cournot duopoly with imperfect substitutes. 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 submodular. The payoffs are chosen to make the best reply mapping simple. 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 and we would lose the symmetry of the game.

are $s_i \in \{-10, ..., 0\}$ and the underlying game payoff of player i reads

$$u_i(s_i, s_j) = \max\{-(\frac{19}{2} + \frac{1}{2}s_i + s_j)s_i, 0\} \tag{17}$$

This transformation makes i's payoff increasing in opponent's action but preserves symmetry, concavity of payoffs {2}, and submodularity {3}. First, increasing player $i's$ action by one unit from s_i increases the payoff of the opponent:

$$u_i(s_i, s_j + 1) - u_i(s_i, s_j) = -s_i > 0 \tag{18}$$

Second, $\delta = -1$, $\sigma = 0$. And third, $\phi = -1$. Thus, all other assumptions hold but is {3} violated.

Condition (16) requires that $i's$ marginal payoff, $-10 - s_j - s_i$, is non-positive if s is agreeable for i. Thus, an agreeable action profile satisfies $m \in \{s|10 + s_j + s_i \geq 0,\ i = 1, 2\}$. Notice, that player i's underlying game best reply to s_j is

$$BR_i(s_j) = -10 - s_j \tag{19}$$

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, $\beta(5, 5) = 0$ as required in theorem 13.

Let's suppose that the guilt cost is as in (2). This guilt cost is supermodular in its arguments and convex in u_i as required in proposition 11. The proof of proposition 9 states that, off the best reply correspondences, a 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

$$10 + s_j + s_i + \theta_i[u_i(s) - 10]s_j \tag{20}$$

This is increasing in a player's own action but the effect of the opponent's action is ambiguous (as opposed to proposition 11 which assumes that the game is supermodular).

So as to the effect of the own action, since $\delta = -1$, $\sigma = 0$ increasing a player's agreed action increases the player's marginal benefit from breaching and leaves the marginal harm unaffected. Within the agreeable set, the agreed payoff effects are as before: thus, the agreed payoff decreases in the player's own action. To sumarize, the marginal incentive to breach is indeed increasing in a player's own action.

Yet, if we consider the effect of the opponent's action, now since the game is submodular, $\phi = -1$, rather than supermodular, increasing the opponent's action decreases the marginal harm on the opponent and decreases a player's marginal benefit from breaching. Agreed payoff increases in a player's own action, as before. The agreed payoff effect and the other two effects now run counter to each other. Thus, the effect on the opponent's incentive to breach is ambiguous: the monotonicity of the marginal incentive to breach in agreed actions (proposition 11) is lost.

Now, let us move on and consider theorem 13 which studies whether efficient agreements can be made, if any. Figure 3 studies the positive quantity equivalent

of the game.[25] There, we suppose that the proneness to guilt is $\theta_i = \frac{1}{7}$ for both players. The action profiles marked with a plus sign are agreeable for player 1 and the action profiles marked with a cross are agreeable for player 2. Thus, the action profiles marked with an asterisk belong to the agreeable set. There are two symmetric action profiles in this set: the equilibrium $(5, 5)$ and $(4, 4)$. Yet, the efficient symmetric action profile $(3, 3)$ (marked with a circle) does not belong to the agreeable set.[26]

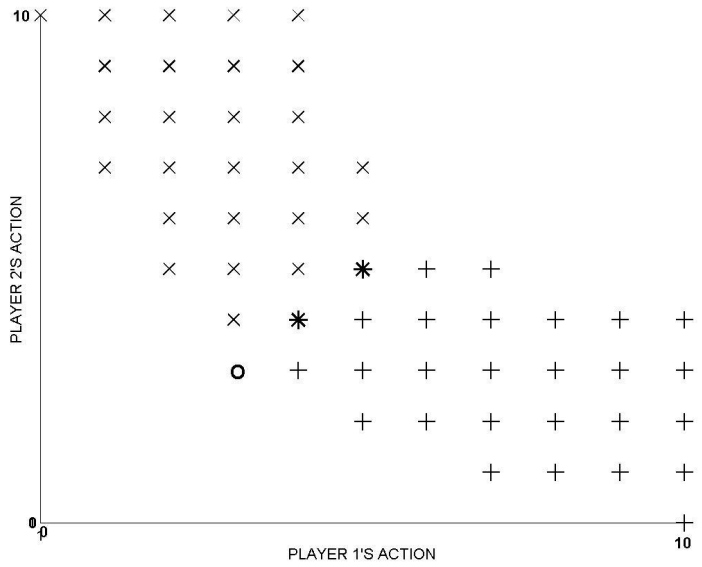

Figure 3: The agreeable set in the Cournot duopoly.

The underlying game equilibrium $(-5, -5)$ is agreeable by proposition 3. To see that $(-4, -4)$ is agreeable, we check that the marginal incentive to breach is negative, $10 - 4 - 4 - \frac{4}{7}[14 - 10] < 0$. 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 $s = (-3, -3)$.

This is because marginal symmetric changes of both actions (i) increase the marginal benefit by $-\delta - \phi = 2$ where both terms are strictly positive, (iib)

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

[26] To see that $(3, 3)$ is efficient, maximize

$$\max_{\sigma}\{-(\frac{19}{2} + \frac{3}{2}\sigma)\sigma\} \tag{21}$$

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 agreed payoffs for the symmetric action profiles 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.

decrease the marginal harm by $\sigma + \phi = -1$, and (iia) change the marginal effect of an increasing agreed payoff by $\delta + 2\phi + \delta = -4 < 0$. The negative marginal effect on the marginal incentive to breach (iia) is vanishing but the positive marginal effects are constant and thus getting relatively stronger as the agreed payoff is increased by symmetric changes of both actions. Thus, even if there is a non-equilibrium action where guilt offsets 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 13.

8 Discussion

The main contribution of this paper is to provide a game theoretic approach to pre-play negotiations and conventions or social norms when people may feel guilty about breaching an agreement. The model incorporates the most important stylized facts that research in social psychology and experimental economics has established about guilt: on the one hand, experiments conducted by Ellingsen and Johanneson (2004), Gneezy (2005) and Charness and Dufwenberg (2003) suggest that people dislike lying and that the more harm they inflict on others by doing so, the more they dislike it. In pre-play negotiations, agreeing to choose an action that one intends not to choose amounts to lying. On the other hand, research in social psychology reveals important stylized facts about guilt that is felt about transgressing social norms in general, not about lying or breaching agreements in particular.

We show that guilt, conventions and pre-play negotiation may have dramatic effects on strategic interaction. Trivially, the set of agreeable outcomes may be larger than the set of underlying game Nash equilibria, since the guilt cost provides an extra incentive to comply to an agreed action profile. Conventions that are not equilibria in the underlying game are possible.

Moreover, the dramatic effects may prevail even if monetary stakes are high: in the prisoner's dilemma, increasing the benefit of defection sufficiently while keeping the harm on the opponent constant will restore the cheap talk prediction that an agreement on cooperation will be breached; yet, no matter how large the benefit of lying, an agreement on cooperation will be credible when the harm that the defection inflicts on the opponent is sufficiently high. Notice also that a player does not become more reluctant to agree on cooperation when she suffers more from defection. Quite the opposite: greater potential harm on herself increases the opponent's relative preference for cooperation since the opponent's promise to cooperate may become credible.

The theory presented is in line with results from public good experiments without contribution thresholds where communication significantly increases contribution levels (Ledyard, 1995). Our theory tells us that for sufficiently high marginal per capita return, the benefit of breaching a cooperative agreement is offset by the harm on opponents. Thus, cooperative agreements become credible[27].

[27] Game theory generally abstracts from where utilities come from. That all agents con-

This result extends to a large class of games with a public good structure: moral hazard in teams, collusion in Bertrand and Cournot duopolies etc. Yet, there is an important distinction as to whether the theory predicts an efficient or an inefficient agreement: in symmetric games with strategic complements, a symmetric non-underlying game equilibrium is agreeable if and only if a symmetric efficient action profile is agreeable. On the other hand, in symmetric games with strategic substitutes, there tends to be a conflict between the incentives to respect an agreement and the efficiency of the agreement: trading off private benefits and the harm on others makes it harder to agree on more efficient actions.

Experiments provide a strong support for our theory. Isaac, McCue and Plott (1985) adopt a decreasing returns to scale production technology for the public good. This implies that actions are strategic substitutes. Despite the positive effect of communication on efficiency, they find that the average contribution levels fall far below the first-best (15% efficient). Isaac and Walker (1988) adopt a constant returns to scale technology implying that actions are weak strategic substitutes. They find a strong positive effect of communication on efficiency. Average contribution levels are up to 99% efficient.

Furthermore, notice that some public good games with a contribution threshold have subsets of the strategy space where actions are strategic substitutes rather than strategic complements. For instance, the stylized version of a public good game with a threshold, the chicken game, has strategic substitutes. In the threshold public good experiments, the effect of communication on contributions has not always turned out to be significant. Thus, even experiments in threshold environments seem to lend support for our theory.

Second, as indicated, our results can be extended to analyze the enforcement power of commonly known conventions and social norms [28]. This is because we abstract from the negotiation protocol and only analyze the interaction when an agreement is in place. Norms here require choosing a particular action in a given situation[29]. In this case, of course, no lies are told per se. Yet, research in social psychology suggests that guilt about transgressing such exogenous norms is stronger the more harm is inflicted on others[30] and, thus, property {A} among

tribute nothing to the public good implicitly assumes that players are money maximizers. Empirical evidence shows that people contribute positive amounts even when the game is played without communication. Distributional preference models map monetary payoff profiles to individual utilities. Thus, the underlying game (in utils payoffs) may differ from a public good game and it may have equilibria where positive amounts are contributed. Even if distributional preferences may be present in the game that is played without communication, guilt combined with pre-play negotiation complements distributional preference motivations in providing the extra incentive to contribute that is in line with what public good experiments have found.

[28] I thank Joel Sobel for poiting this out.

[29] Social norms can be considered to be established by a community's moral discourse -- grand scale pre-play negotiations: When John Doe violates a social norm, the violation launches a vivid discourse by others in the community. This discussion may involve arguments for and against John Doe's action. If the social norm is well established arguments are mostly against and parties quickly converge into an agreement on how John should have behaved.

[30] For evidence, see related literature in psychology in section 2.

others remains valid. Thus, the theory can be interpreted as a tractable model of let-down aversion where pre-play negotiation or convetions establish commonly known, coinciding and justified mutual expectations about behavior.

The theory presented in this paper has a further interpretations in addition to face-to-face communication and convetions in one-shot games. Analogous results to those presented in this paper would be obtained if we suppose that players have zero proneness to guilt and they informally agree on a stationary outcome in an infinitely (and infinitely often) repeated analog of the underlying game. 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 deviator's agreed payoff and the harm she inflicts 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 internalized.[31]

Pre-play communication may have other functions than implementing shared standards of behavior. Pre-play communication could improve upon players' knowledge and degree of knowledge about the relevant game and thus indirectly promote certain actions as ideals of behavior. Our model abstracts from these functions. Also, these functions may be relevant if we try to model how a community communicates to reach a common understanding of its environment.

This paper has not analyzed the effect of the negotiation protocol on the agreement. A cooperative solution concept or a bargaining protocol can be applied in predicting which agreement will be chosen from the set of agreeable profiles. When proneness to guilt is zero, the smallest agreeable set is the set of (non-cooperative) Nash equilibria of the UG. When the proneness to guilt is infinite any agreement that no player can unilaterally Pareto-improve upon is agreeable. Thus, as we increase the players' proneness to guilt from zero to infinite, we move from an entirely non-cooperative prediction to a largely cooperative one. We intend to study the effect of the negotiation protocol in a follow-up paper which also presents experimental evidence of such effects.

Another dimension for future research is the relaxation of the assumption of complete information of proneness to guilt types. The choice of an optimal agreement when information is private requires trading off the own agreed payoff with the probability that the opponent breaches the agreement.[32] On the other hand, a dynamic setup of incomplete information on proneness to guilt would allow for the players to build up reputations. First, it may be 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 may be willing to build up a reputation for

[31] See appendix A for further details.

[32] Notice yet, that if the information on proneness to guilt is private, signalling is not an issue: the maximisation problem conditional on respecting is the same independently of the type, and thus all types that intend to respect behave identically. Any type who intends to breach is thus detected. Thus, her opponent knows that she will not suffer guilt.

a higher proneness to guilt in order to be able to reach agreements with a larger fraction of types. From a similar perspective one can study the evolution of proneness to guilt for a given stochastic process of games and matches.

9 Appendix

9.1 A) Repeated games

Results analogous to those in this paper would be obtained, 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 internalized 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 identical to that in the single shot model, 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 $k(h_j(m, a_i), v_i(m), u_i^P)$ takes the following form

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

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

9.2 B) Exogenous action-norms and moral discourse

Harsanyi (1977) and Binmore (1998) present models where a social contract is agreed upon in a moral discourse which is considered to take place prior to the play of the grand game of life[33]. The social contract can be interpreted as a collection of action-norms that apply in various circumstances in the grand game of life. As indicated by psychological research, violating such norms causes distress, such as guilt, shame and fear of punishment. Another stylized fact

[33] Similar philosophical non-game theoretic approaches are provided by Habermas (1990) and Hoppe (1993), for instance.

of the research in psychology is that guilt (or distress) is proportional to the harm that violation causes on others. Thus, the approach developed here can be applied to general action-norms as outcomes of moral discourse - pre-play negotiations of the grand game of life.

In the game theoretic models of Harsanyi (1977) and Binmore (1998), players have empathetic preferences which are weighted sums of individual preferences and used in moral discourse to derive a shared perception of a fair social contract. The fairness preferences are derived from weighting of the individual preferences in an impartial original 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 action profiles of play and $\{1,2\}$ is the set of possible roles. A 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 2. The formulation $U_i(m,s) = u_i(s) + \theta_i v_i(m) h_j(m,s)$ is reached by letting the weight depend on the agreed payoff $v_i(m)$. The implication is thus a truncated additive social welfare function where the concern for the opponent depends on how nicely one is treated in the pre-play negotiations and how prone to guilt (empathetic) one is.

9.3 C) Proofs

9.3.1 Proof of lemma 2

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

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

Suppose now that m is agreeable. But, then by definition (IC_i) holds.■

9.3.2 Proof of proposition 3

Since m is an equilibrium $v_i(m) \geq 0$ for $i = 1,2$. Since m is an equilibrium in Γ, , $m_i \in BR_i(m_j)$ for $i = 1,2$. Then, by lemma (2), $m \in A_i(\Gamma,\theta_i)$ for $i = 1,2$ and, by definition, $m \in A(\Gamma,\theta)$.■

9.3.3 Proof of lemma 4

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

9.3.4 Proof of lemma 5

Conditions (AC) and (EF) imply that $g_i(v_i(m), h_j(m, s_i)) = 0$ if $h_j(m, s_i) < 0$. But indeed, $h_j(m, s_i) \doteq u_j(m) - u_j(m_j, s_i) \leq 0$. Thus (IC_i) is violated and $m \notin A_i(\Gamma, \theta_i)$ and thus $m \notin A(\Gamma, \theta)$.■

9.3.5 Proof of proposition 6

By assumption, $v_i(m) > 0$ for $i = 1, 2$. Take player i and an arbitrary s_i. First, if $u_i(m_i, m_j) \geq u_i(s_i, m_j)$ then $u_i(m_i, m_j) \geq u_i(s_i, m_j) - g(v_i(m), h_j(m, s_j))$ and $B_i(m, s_i; \theta_i) \leq 0$. Second, if $u_j(m_j, s_i) < u_j(m_j, m_i)$ then $h_j(m, s_i) > 0$. By, (EF) $g(v_i(m), h_j(m, s_i)) > 0$. Thus, since payoffs in Γ are finite, $\lim_{\theta_i \to \infty} \theta_i g(v_i(m), h_j(m, s_i)) \geq u_i(s_i, m_j) - u_i(m_i, m_j)$. Hence, $\lim_{\theta_i \to \infty} B(m, s_i; \theta_i) \leq 0$. Since either $u_i(m_i, m_j) \geq u_i(s_i, m_j)$ or $u_j(m_j, s_i) < u_j(m_j, m_i)$ holds for every s_i, (IC_i) holds. Thus $m \in A_i(\Gamma, \theta_i)$. This is true for both players. Thus, $m \in A(\Gamma, \theta)$.

Let now $m \in A(\Gamma, \theta)$. Suppose to the contrary that there is i and s_i such that neither $u_i(m_i, m_j) \geq u_i(s_i, m_j)$ nor $u_j(m_j, s_i) < u_j(m_j, m_i)$ holds. Then, both are true. But then, by lemma 5, $m \notin A(\Gamma, \theta)$. This is a contradiction.■

9.3.6 Proof of proposition 9

Lemma 5

Lemma 14 *Let Γ be finite. Let $m_i \neq \{s_i^1, s_i^n\}$. Let {1} , {2} and {4} hold. Then (IC_i) holds if and only if $\mathbb{B}_i(m, \theta_i) \leq 0$.*

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

Let $\mathbb{B}_i(m, \theta_i) > 0$. If $u_i(m_i - 1, m_j) - u_i(m_i, m_j) \geq 0$, $B(m, m_i - 1; \theta_i) > 0$ and (IC_i) is violated. If $u_i(m_i + 1, m_j) - u_i(m_i, m_j) > 0$, then $B(m, m_i + 1; \theta_i) > 0$ and (IC_i) is violated.

Let (IC_i) be violated. Thus, there is s_i' such that $B_i(m, s_i'; \theta_i) > 0$. Suppose to the contrary that $\mathbb{B}_i(m, \theta_i) \leq 0$. We only need to consider the case $u_i(m_i - 1, m_j) - u_i(m_i, m_j) \geq 0$ since if $u_i(m_i + 1, m_j) - u_i(m_i, m_j) > 0$, then $\mathbb{B}_i(m, \theta_i) > 0$ by definition.

Let thus $u_i(m_i - 1, m_j) - u_i(m_i, m_j) \geq 0$. By assumption $\mathbb{B}_i \leq 0$ and thus

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

By assumption {1}, the harm increases in deviations further downwards. Also by assumption {4} guilt cost is convex in h_j and by assumption {2} u_j is concave in s_i. Thus the harm is convex in s_i and the guilt cost is also convex in s_i as a composite of two convex functions. 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 $\mathbb{B}_i(m, \theta_i) \leq 0$ then $B(m, s; \theta_i) \leq 0$ for all $s_i < m_i$. We have a contradiction. ■

Proof of the proposition The result follows directly from lemma 2, lemma 14 and the fact that $A(\Gamma, \theta) = \cap_{i=1,2} A_i(\Gamma, \theta_i)$■

9.3.7 Proof of lemma 10

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

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

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

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

9.3.8 Proof of proposition 11

Since u_i is increasing in s_j, by lemma 5, 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

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

But $\beta_i(m)$ is increasing in m_i and $\eta_j(m)$ is decreasing in m_i by lemma 10. Also, $u_i(m_i,m_j) - u_i(m_i-1,m_j) < 0$ implies that $u_i(m)$ decreases in m_i. But g is increasing in both arguments. Thus, $\mathbb{B}_i(m_i,m_j)$ is indeed increasing in m_i.

On the other hand, $\beta_i(m_i,m_j)$ is decreasing in m_j and $\eta_j(m_i,m_j)$ is increasing in m_j by lemma 10. Also, u_i is increasing in m_j by assumption. But g is increasing in both arguments. Thus, $\mathbb{B}_i(m_i,m_j)$ is indeed decreasing in m_i.■

9.3.9 Proof of theorem 12

Since $\phi_i(s) < 0$ for all s, the best reply curves are downward sloping. Since payoff functions are concave, for $s_i < BR_i(s_j)$, $u_i(s_i+1,s_j) - u_i(s_i,s_j) > 0$. For any symmetric action profile such that $s_i < s_i^*$, $s_i < BR_i(s_j)$. Thus increasing the action of i improves the payoff of both. Thus symmetric profiles such that $s_i < s_i^*$ are not efficient. Thus, If $u_i(s_i^*+k, s_j^*+k) > 0$ for some $k \in \mathbb{Z}$ then $k > 0$. But, since $\phi_i(s) < 0$, by lemma 10, $\beta_i(s^*+k) > \beta_i(s^*)$ and $\eta(s^*+k) < \eta(s^*)$ for $i = 1,2$.■

9.3.10 Proof of theorem 13

Lemmas 15 to 20

Lemma 15 *If $\delta \neq \phi$ then there is at most one equilibrium s^* where $\beta_i(s^*) = 0$ for $i = 1,2$*

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

If $-\delta < \phi$ then $(s_i,s_j) \in M_F$ implies $s_i \leq s^*$ 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, $s^* = (s_1^*, s_2^*)$. $s \in M_F$ implies that $\frac{\partial u_i(s)}{\partial s_i} \leq 0$ for $i = 1,2$.

For player two this is true for $m_2 \geq BR_2(m_1)$ and for player one this is true for $m_2 \geq BR_1^{-1}(m_1)$. Thus the claim. ■

Lemma 16 *Let the game be symmetric. If s^c maximizes $max_{k\in Z}u(s+k)$ 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 $u_i(s^c) > u_i(s^c + k) = u_i(s_i', s_i') > u_i(s_i', s_j')$ 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 $k \in \mathbb{Z}$, there must be k that maximizes $u(s + k)$ with $s_i = s_j$ and $s \in S$. ■

Lemma 17 *$s \notin M_F \Rightarrow$ there is i such that $u_i(s + 1) > u_i(s)$*

Proof. $s \notin M_F \Rightarrow$ there is i such that $u_i(s + 1) - u_i(s) = [u_i(s + 1, s + 1) - u_i(s, s + 1)] + [u_i(s, s + 1) - u_i(s, s)] > 0$. ■

Lemma 18 *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. ■

Lemma 19 *Let $\sigma + \delta < 0$, $2\phi + \delta + \sigma \geq 0$ and $\phi \geq 0, \delta \leq 0, \sigma \leq 0$. Let $u_i(s) - u_i(s-1) \geq 0$ and $\beta_i(s-1) \geq 0$. Let g satisfy {4} and {5}. Suppose that $\beta_i(s-1) \geq g(u_i(s-1), \eta_j(s-1))$. If $\beta_i(s) \leq g(u_i(s), \eta_j(s))$ then $\beta_i(s + k) \leq g(u_i(s + k), \eta_j(s + k))$ for all $k > 0$.*

Proof. $\delta + 2\phi + \sigma \geq 0$ and $\phi + \delta < 0$ implies that $\phi + \sigma \geq 0$. Then, by lemma 10, $\beta(s+k)$ is increasing and concave in k and $\eta(s+k)$ is increasing and convex in k.

Since $\delta + 2\phi + \sigma \geq 0$ and $u_i(s) - u_i(s - 1) \geq 0$, $u(s + k)$ is convex and increasing in k for $k \geq 0$. Thus, $g(u(s + k), \eta(s))$ is convex and increasing in k since g is convex and increasing in u by $\{5\}$. Similarly, $g(u(s), \eta(s + k))$ is convex and increasing in k since g is convex in η for $\eta \geq 0$ by $\{4\}$.

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

$$\begin{aligned} & \beta(s) - \beta(s-1) \\ \leq \; & g(u(s), \eta(s)) - g(u(s-1), \eta(s-1)) \end{aligned}$$

Thus, by lemma 18 and since g is supermodular in its arguments

$$\begin{aligned} 0 &\leq \beta(s+1) - \beta(s) \\ &= -\delta - \phi \\ &= \beta(s) - \beta(s-1) \\ &\leq g(u(s), \eta(s)) - g(u(s-1), \eta(s-1)) \\ &\leq g(u(s+1), \eta(s+1)) - g(u(s), \eta(s)) \end{aligned}$$

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

Lemma 20 *Let $\sigma + \delta < 0$, $\phi + \sigma \geq 0$ and $\phi \geq 0, \delta \leq 0, \sigma \leq 0$. Let $\beta_i(s-1) \geq 0$. Let g satisfy {4} and let $g(u', \eta) = g(u, \eta)$ for all u', u, η. Suppose that $\beta_i(s-1) \geq g(u_i(s-1), \eta_j(s-1))$. If $\beta_i(s) \leq g(u_i(s), \eta_j(s))$ then $\beta_i(s+k) \leq g(u_i(s+k), \eta_j(s+k))$ for all $k > 0$.*

Proof. By lemma 10, $\beta(s+k)$ is increasing and concave in k and $\eta(s+k)$ is increasing and convex in k.

Also $g(u(s), \eta(s+k))$ is convex and increasing in k since g is convex in η for $\eta \geq 0$ by $\{4\}$ and for all u, $g(u, \eta(s+k)) = g(u(s+k), \eta(s+k))$ by assumption.

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

$$\begin{aligned} & \beta(s) - \beta(s-1) \\ \leq \; & g(u(s), \eta(s)) - g(u(s-1), \eta(s-1)) \end{aligned}$$

Thus, since $g(u(s), \eta(s+k))$ is convex and increasing in k

$$\begin{aligned} 0 &\leq \beta(s+1) - \beta(s) \\ &= -\delta - \phi \\ &= \beta(s) - \beta(s-1) \\ &\leq g(u(s), \eta(s)) - g(u(s-1), \eta(s-1)) \\ &= g(u(s+1), \eta(s)) - g(u(s), \eta(s-1)) \\ &\leq g(u(s+1), \eta(s+1)) - g(u(s), \eta(s)) \end{aligned}$$

We can proceed by induction to show that for every $s+k$ with $k > 0$, we have $\beta(s+k) - g(u(s+k), \eta(s+k)) \leq \beta(s) - g(u(s), \eta(s)) \leq 0$. Thus every $s+k$ with $k > 0$ is agreeable.

Proposition 21 *Let $\delta + 2\phi + \sigma \geq 0$. Let $s' - 1 \in M_F$ and $s' - 1 \notin A(\Gamma, \theta)$. Let $u(s') - u(s' - 1) \geq 0$. Let s^* be the unique equilibrium of the game and $\beta_i(s^*) = 0$. Suppose that {1}, {2}, and {3} hold. Furthermore, let g satisfy {4}, and {5}. If s' such that $s'_i > s^*_i$ is agreeable then any $s' + k$ such that $k > 0$ is agreeable.*

■

Proof. $\delta + 2\phi + \sigma \geq 0$ implies that $\phi + \delta \geq 0$ or $\sigma + \phi \geq \phi$. Suppose first that $\delta + \phi > 0$. By lemma 15 $s'_i - 1 < s^*_i$ for $i = 1, 2$. Thus $s'_i \leq s^*_i$ and the claim holds trivially.

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

Let now, $\phi + \delta < 0$. Then $\sigma + \phi > 0$. The fact that $s' - 1 \in M_F$ implies that $\beta(s' - 1) \geq 0$. By lemma 10, $\beta(s + 1) - \beta(s) = -\delta - \phi > 0$. Also, since $\phi > -\sigma$, by lemma 10 $\eta(s + 1) - \eta(s) = \sigma + \phi > 0$ and thus $\eta(s + k)$ is weakly increasing in k.

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

Since $u(s + k)$ is convex in k, then $g(u(s + k), \eta(s))$ is convex in k since g is convex in u. Similarly, $g(u(s), \eta(s + k))$ is convex and increasing in k for $k \geq 0$ since g is convex and increasing in η for $\eta \geq 0$ and $\eta(s + k)$ is convex and increasing in k for $k \geq 0$.

Also $s' - 1 \in M_F$ and since $s' - 1 \notin A(\Gamma, \theta)$ we have $\beta_i(s' - 1) > 0$ and $\beta_i(s' - 1) - g(u_i(s' - 1), \eta_j(s' - 1)) > 0$. But $s' \in A(\Gamma, \theta)$, and $\beta_i(s') \geq \beta_i(s' - 1) > 0$. Thus, by lemma 19 every, $s' + k$ with $k > 0$ is agreeable. ■

Proof of the theorem By proposition 3, s^* is agreeable as a Nash equilibrium of the underlying game. Thus '$\Longrightarrow$' is trivial.

Let us now show that if a symmetric $s \neq s^*$ is agreeable,then an efficient symmetric s is agreeable.

If $\phi + \delta \geq 0$ then since the game is symmetric and there is an inefficient equilibrium such that $\beta_i(s^*) = 0$, (s^n_1, s^n_2) is an equilibrium. To see this consider two subcases, 1) $\phi + \delta > 0$ and 2) $\phi + \delta = 0$. If $\phi + \delta > 0$, since $\phi > 0$, the best reply correspondences are upward sloping and steeper than one and they cross at s^*. By symmetry (s^n_1, s^n_2) is an equilibrium, since $-\beta_i(s^n) > 0$ for $i = 1, 2$. If $\phi + \delta = 0$ and there is s^* such that $\beta_i(s^*) = 0$ for $i = 1, 2$, then both best reply functions have a slope equal to one and they overlap in the entire strategy space. All symmetric profiles are equilibria and thus (s^n_1, s^n_2) is an equilibrium. In both cases, since u_i is increasing in s_j, by theorem 7 in Milgrom, Roberts (1990), (s^n_1, s^n_2) is efficient and by proposition 3, (s^n_1, s^n_2) is agreeable.

Suppose now that $\phi + \delta < 0$. Then if $\sigma + \delta + 2\phi \geq 0$, we must have $\sigma + \phi > 0$. Since $\sigma + \delta + 2\phi \geq 0$ and $\beta_i(s^*) = 0$ for $i = 1, 2$, (s^n_1, s^n_2) is efficient. To see this, first consider profiles $s^* - k$ for $k > 0$. By lemma 15, $s^* - k \notin M_F$. By lemma 17, there is i such that $u_i(s^* - k + 1) - u_i(s^* - k) > 0$. Since the game is symmetric, this holds for both players. But since $\sigma + \delta + 2\phi \geq 0$, $u_i(s + k)$ is convex in k. Thus, $u_i(s + k) > u_i(s)$ for $i = 1, 2$ for all symmetric s and for all

$k > 0$. Thus, (s_1^n, s_2^n) maximizes the payoff along the diagonal. Thus, by lemma 16, (s_1^n, s_2^n) is efficient.

Let $\phi + \delta < 0$ still hold and suppose alternatively that $\sigma + \delta + 2\phi < 0$. Then $u_i(s+k)$ is strictly concave in k. By lemma 17 and by symmetry, $u_i(s^*) > u_i(s^*-1)$ for $i = 1, 2$. Since the strategy set is bounded a maximizer s^*+k along the diagonal exists and it satisfies $k > 0$.

Since asymmetric $s \neq s^*$ is agreeable, by lemma 15, $s = s^* + k$ for some $k > 0$. For each player, consider two subcases, 1) there is $1 < k' < k$ such that $s^*+k'-1 \notin A_i(\Gamma, \theta_i)$ but $s^*+k' \in A(\Gamma, \theta)$ and 2) s^*+k' where $k' = 1$ is agreeable. It is easy to see that one of the two must hold for each player. In either case the agreeability of $s^* + k'$ implies that $\beta_i(s^* + k') \leq g_i(u(s^* + k'), \eta_j(s^* + k'))$ and in each subcase $\beta_i(s^* + k' - 1) \geq 0$ and $\beta_i(s^* + k' - 1) \geq g_i(u(s^* + k' - 1), \eta_j(s^* + k' - 1))$. Thus, if $\sigma + \delta + 2\phi \geq 0$ and g satisfies $\{5\}$ we can apply lemma 19. On the other hand, if $\phi + \sigma \geq 0$ and $g(u', \eta) = g(u, \eta)$ for all u', u, η, we can apply lemma 20. In either case any $s^* + k$ with $k \geq k'$ is agreeable. Thus an efficient symmetric profile is agreeable.■

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, 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. (1998): Game Theory and The Social Contract: Just Playing. MIT Press. Cambridge, MA.

[8] Bohnet, I.; Frey B.S. (1998): The Sound of Silence in Prisoner's Dilemma and Dictator Games. Journal of Economic Behavior and Organization 38, 43-57.

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

[10] 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.

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

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

[13] Crawford, V.P. (2003): Lying for Strategic Advantage: Rational and Boundedly Rational Mispresentations. American Economic Review 93, 133-149.

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

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

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

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

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

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

[20] 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.

[21] Duffy, J.; Feltowich N. (2004): Words, Deeds and Lies: Strategic Behavior in Games with Multiple Signals. Forthcoming in the Review of Economic Studies.

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

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

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

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

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

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

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

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

[30] Habermas, J. (1984): The Theory of Communicative Action. Beacon Press, Boston.

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

[32] 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.

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

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

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

[36] 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.

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

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

[39] Miettinen, T. (2005): What is Guilt Like. On the Concept of Guilt Aversion. Mimeo.

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

[41] Pavitt, C.; Shankar, A. (2002): Resource and Public Good Dilemmas: A New Issue for Communication Research. The Review of Communication. 2, 251-272.

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

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

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

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

[46] Sobel, J. (2005): Interdependent Preferences and Reciprocity. Journal of Economic Literature. 43, 2.

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

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