Demichelis, Stefano; Weibull, Jörgen W.

Working Paper

Efficiency, communication and honesty

SSE/EFI Working Paper Series in Economics and Finance, No. 645

Provided in Cooperation with:
EFI - The Economic Research Institute, Stockholm School of Economics

Suggested Citation: Demichelis, Stefano; Weibull, Jörgen W. (2006) : Efficiency, communication and honesty, SSE/EFI Working Paper Series in Economics and Finance, No. 645, Stockholm School of Economics, The Economic Research Institute (EFI), Stockholm

This Version is available at:
https://hdl.handle.net/10419/56079

EFFICIENCY, COMMUNICATION AND HONESTY

Stefano Demichelis* and Jörgen W. Weibull†

SSE/EFI Workning Paper Series in Economics and Finance
No. 645

First draft February 1, 2006. This version: November 28, 2006.

Abstract. We here develop a model of pre-play communication that generalizes the cheap-talk approach by allowing players to have a lexicographic preference, second to the payoffs in the underlying game, for honesty. We formalize this by way of an honesty (or truth) correspondence between actions and statements, and postulate two axioms met by natural languages. The model is applied to finite and symmetric two-player games and we establish that honest communication and play of the Pareto dominant Nash equilibrium together characterize the unique evolutionarily stable set in generic and symmetric $n \times n$-coordination games. In particular, this holds even in Aumann's (1990) example of a Pareto dominant equilibrium that is not self-enforcing.

Keywords: Efficiency, communication, coordination, honesty, evolutionary stability.

JEL-codes: C72, C73, D01.

1. Introduction

Communication is crucial to most human interaction, and yet traditional economic analyses either neglect communication or presume that it leads to play of an equilibrium that is not Pareto dominated by another equilibrium.[1] An example of the latter is renegotiation proofness, a criterion that is sometimes invoked in contract theory and in analyses of repeated games (see Benoit and Krishna (1993) for a succinct

*Department of Mathematics, University of Pavia, Italy. Demichelis thanks the Knut and Alice Wallenberg Foundation for financial support and the Stockholm School of Economics for its hospitality. Both authors thank Cedric Argenton, Milo Bianchi, Vince Crawford, Segismundo Izquierdo and Robert Östling for comments.

†Department of Economics, Stockholm School of Economics.

[1]Laboratory experiments generally support the hypothesis that pre-play communication leads to play of such equilibria in coordination games. A pioneering study of this phenomenon is Cooper et al. (1989). See Crawford (1998) for a survey, Charness (2000), Clark, Kay and Sefton (2001) and Blume and Ortmant (2005) for more recent contributions.

analysis). However, as pointed out by Aumann (1990), strategically interacting decision makers may agree to play a Pareto dominant equilibrium even if each decision maker secretly plans to deviate. Aumann illustrated this possibility by means of the following two-player game:

$$\begin{array}{cc|cc} & & c & d \\ \hline c & & 9,9 & 0,8 \\ d & & 8,0 & 7,7 \end{array} . \tag{1}$$

This game has three Nash equilibria, all symmetric: the Pareto dominant but risk dominated strict equilibrium (c, c), the risk dominant but Pareto dominated strict equilibrium (d, d), and a mixed equilibrium that results in an intermediate expected payoff. Aumann points out that each player has an incentive to suggest play of (c, c), even if the suggesting player actually plans to play d; it is advantageous to make the other play c rather than d irrespective of what action the suggesting player takes. In Aumann's colorful words, with Alice and Bob in the two player roles:

> 'Suppose that Alice is a careful, prudent person, and in the absence of an agreement, would play d. Suppose now that the players agree on (c, c), and each retires to his "corner" in order actually to make a choice. Alice is about to choose c when she says to herself: 'Wait; I have a few minutes; let me think this over. Suppose that Bob doesn't trust me, and so will play d in spite of our agreement. Then he would still want me to play c, because that way he will get 8 rather than 7. And of course, also if he does play c, it is better for him that I play c. Thus he wants me to play c no matter what. [...] Since he can reason in the same way as me, neither one of us gets any information from the agreement; it is as if there were no agreement. So I will choose now what I would have chosen without an agreement, namely d.'" (op. cit. p. 202)

Aumann concludes that the strict and Pareto dominant Nash equilibrium (c, c) is not self-enforcing. This line of reasoning abstracts away from the possibility that Alice and Bob have a preference for honesty (here: for not deviating from an agreement). In this abstraction, Aumann is not alone. Indeed, virtually all of modern economics relies on the presumption that economic agents have no preference for honesty, or more specifically against lying, *per se*. The standard assumption in all of information economics (ranging from mechanism design to the market for lemons) is that economic agents misreport their private information whenever they believe it is in their interest

to do so.[2]

The purpose of the present study is to investigate the implications of a (weak) preference for honesty in coordination games with a pre-play communication stage. We show that a lexicographic preference for honesty (second to the payoffs in the underlying game) rules out, in the long run, behaviors such as the one described in the above example, and, more generally, implies honest communication and play of the Pareto dominant equilibrium in all symmetric $n \times n$-coordination games with a unique Pareto dominant equilibrium. We achieve this by way of a generalization of the cheap-talk approach to include what we call an *honesty* (or, somewhat more narrowly, *truth*) *correspondence*, a correspondence that specifies what pre-play messages are honest (true) when uttered in conjunction with a given action in the underlying game G. For instance, the statement "I will play c" is honest (and true) if and only if the player actually takes action c (in the absence of any risk that intentions cannot be carried out) while "I will play c or d" is honest (true), irrespective of what action the speaker takes. Statements such as "I suggest that we play (c, c)" or "Let us agree to play (c, c)", if followed by play of d, are neither true nor false, but would be deemed dishonest, we believe, by most people.[3]

Most individuals presumably feel some guilt or shame when lying or being dishonest. Gneezy (2005) provides experimental evidence for a psychological cost associated with the act of lying (see also Ellingsen and Johannesson (2004) and Hurkens and Kartik (2006)). Gneezy's main empirical finding is that "...people not only care about their own gain from lying; they also are sensitive to the harm that lying may cause the other side. The average person prefers not to lie, when doing so only increases her payoff a little but reduces the other's payoff a great deal." (op. cit. p. 385). In the context of the above example: for a sufficiently large psychological cost of lying, neither Alice nor Bob would say that they will play c and then play d. What happens, by contrast, if the preference for honesty is weak? This is exactly what we analyze here. We go to the extreme and assume that players have a lexicographic preference for honesty, that they avoid dishonest statements only if this comes at *no* loss of material payoffs. This assumption may, at first sight, seem too weak to have any interesting implication for behavior. However, this is not so. For example, suppose that, in Aumann's example, both Alice and Bob first say that they will play c and yet they each then takes action d. Such behavior is compatible with Nash equilibrium

[2]Notable exceptions are Alger and Ma (2003) and Alger and Renault (2006a,b).

[3]Examples of lying that is usually not thought to be dishonest are "white lies" in social life and policy makers' denials of plans to devalue a currency.

under cheap talk, since then messages have no exogenous meaning. By contrast, it is incompatible with Nash equilibrium in our lexicographic communication game if the language is rich enough. For if the language contains some message, m, that is honest only if action c is taken and another message, m', that is honest only when followed by action d — two innocuous assumptions about any natural language — then it it would be lexicographically better to say m' instead of m, since there can follow no payoff loss in G.[4]

This is not the end of the analysis, however. First, it is easily shown that there are Nash equilibria in the lexicographic communication game in which both players are dishonest. Secondly, there are Nash equilibria that are Pareto dominated by other Nash equilibria. However, we establish that set-wise evolutionary stability implies Pareto dominant Nash equilibrium play in finite and symmetric two-player $n \times n$-coordination games, granted the language satisfies two axioms—a precision axiom and a null axiom—axioms that generalize the richness properties alluded to above. We believe that set-valued evolutionary stability is relevant in the present context. For if games are played over and over in a large population with a common language, then drift may occur in continua of payoff-equivalent strategies. Thus, the population state will eventually leave such a continuum, unless strategies in the continuum "defeat" strategies outside the set, which is, roughly, what set-wise evolutionary stability requires. Drift in equilibrium components of games has been analyzed before, see in particular Binmore and Samuelson (1994, 1997).

The mechanism that drives home our efficiency result is similar to that in Robson (1990) in that it depends on the existence of messages that are not sent in equilibrium. Robson noted that, in a population playing such an equilibrium, deviating players can use such messages (as a "secret handshake") to recognize each other and coordinate their play. However, while the existence of such unused messages is presumed in Robson (1990), and non-deviating players in his setting by assumption do not react to such messages, the existence of unsent messages is here derived from primitives and non-deviators may recognize, and hence also punish, senders of such messages.

We believe that our approach is original. It clearly differs from the cheap-talk literature and it also differs from other models of pre-play communication in which messages have a pre-existing meaning. Let us briefly comment on some contributions along the latter line. Farrell (1988,1993) analyzes cheap-talk pre-play communication when messages have a pre-existing meaning. In the second paper, he

[4] Just as with Aumann's informal reasoning, this hinges on the fact that the off-diagonal payoff 8 is no less than the on-diagonal payoff, 7.

defines *neologism-proofness*, a refinement of perfect Bayesian equilibrium in cheap-talk games.[5] Unlike here, players have no preference for or against honesty *per se*. Instead, Farrell imposes a credibility condition on unsent messages, roughly requiring the listener to believe the speaker, unless the speaker has a "good reason" to mislead the listener. Thus, our approach is quite distinct from that of Farrell. Myerson (1989) focuses on the determination of a single negotiation statement made by one individual, the negotiator, while we here focus on the determination of pairs of statements, made by both players in two-player games. Rabin (1994) analyzes two-sided pre-play communication in symmetric two-player games. He considers cheap talk in a language with pre-existing meaning, and players make repeated simultaneous statements before they play the underlying game. Repetition allows them to make agreements to play particular equilibria of the underlying game: if the players propose the same equilibrium in a given pre-play communication round, then this defines an agreement to play that equilibrium. Rabin's modelling approach is clearly different from the one taken here. In particular, we do not presume that "agreements" will be followed but instead presume a lexicographic preference for "keeping one's word." Blume (1998) analyzes stochastic population learning in pre-play communication games where some messages have *a priori* information content, modelled as follows. For each strict equilibrium in the underlying game, each player has exactly one message linked to that equilibrium. If such a linked message is sent, then the receiver of the message obtains a small increase in his or her material payoff if playing his or her strategy in that equilibrium, while the sender's payoff is unaffected. By contrast, we assume that a sender who sends a dishonest message will make a lexicographic payoff loss, while there is no direct effect on the payoff to the receiver of such a message. Crawford (2003) analyzes one-way pre-play communication in zero-sum games in which the players send messages to each other about their intentions, in a pre-existing language, as here. Players do not have a preference against lying *per se*, however, but may be either *sophisticated* or *mortal*. The first category is essentially the usual *homo oeconomicus,* as portrayed in game theory, while representatives of the second category do not always have correct beliefs about their opponent's behavior (in particular, they expect their own attempts to deceive their opponent to always be successful). The possibility of a mortal opponent fundamentally alters the game from sophisticated players' viewpoint, and makes deception possible in sequential equilib-

[5]Farrell (1988) investigates another solution concept for pre-play communication games. a concept that combines elements of Nash equilibrium with elements of rationalizability.

rium.[6] Miettinen (2006), finally, develops a model of pre-play negotiations in which players incur a psychological costs (of guilt) if they breach an agreement and thereby harm the other party. The cost is weakly increasing in the agreed payoff and weakly decreasing in the harm caused the other player. Hence, while addressing a related question, the approach is quite distinct. In particular, it is not a generalization of cheap talk.

The rest of the paper is organized as follows. The model is specified in Section 2, Nash equilibrium is analyzed in Section 3 and evolutionary stability in Section 4. Section 5 concludes.

2. The model

Let G be a symmetric $n \times n$ two-player game with payoff matrix $\Pi = (\pi(i,j))$. Thus, $\pi(i,j)$ is the payoff to a player who uses pure strategy i when the other player uses strategy j. We will refer to G as the *underlying* game. Let A denote the finite set of pure strategies of G, to be called *actions*.

Let M be a non-empty finite set of *messages*, and let $\mathcal{G} = (S, v)$ be a symmetric *cheap-talk communication game*, based on the game G, as follows: first, the players simultaneously send a message to each other, then each player observes the other's message and takes an action in G. The pure-strategy set for each player in $\mathcal{G}$ is thus the finite set

$$S = \{(m, f) : m \in M \ \& \ f : M \longrightarrow A\}, \tag{2}$$

where $m \in M$ is a message to send and f maps the other player's message, m', to an own action $a = f(m')$. Given a mixed strategy $\sigma \in \Delta(S)$, where $\Delta(S)$ is the unit simplex of probability distributions over S, let $\sigma(m, f)$ denote the probability assigned to the pure strategy $s = (m, f)$, and let $M(\sigma) \subset M$ be the set of messages used with positive probability in σ. Let $\sigma^m(m') \in \Delta(A)$ denote the conditional probability distribution induced by σ, conditional on having sent m and received m'. Define $v : S^2 \to \mathbb{R}$ by

$$v[(m, f), (m', g)] = \pi[f(m'), g(m)]. \tag{3}$$

This is the *payoff* in $\mathcal{G}$ to a player who uses pure strategy (m, f) against (m', g). The payoff function v is linearly extended to mixed strategies in $\mathcal{G}$ as usual.

Let $\beta : \Delta(S) \rightrightarrows \Delta(S)$ be the best-reply correspondence in $\mathcal{G}$ and let

$$\Delta^{NE} = \{\sigma \in \Delta(S) : \sigma \in \beta(\sigma)\} \tag{4}$$

[6] For other equilibrium analyses of deceit and lying, see Sobel (1985), Benabou and Laroque (1992), Farrell and Gibbons (1989) and Conlisk (2001).

be the set of strategies in symmetric Nash equilibria of $\mathcal{G}$.

Let $\tilde{\mathcal{G}}$ be a *lexicographic communication game*, derived from $\mathcal{G}$ as follows. The messages, actions and strategies are defined as in $\mathcal{G}$. Call the payoffs in $\mathcal{G}$ *material payoffs*. We proceed to define $\tilde{\mathcal{G}}$ as an *ordinal game*, that is, a game in which players have complete and transitive preference orderings over mixed-strategy profiles (see Chapter 2 in Osborne and Rubinstein (1994)).

All messages in $\tilde{\mathcal{G}}$ have a pre-determined meaning in the sense that for each message $m \in M$ there is a non-empty subset $H(m) \subset A$ of actions such that m is *honest* if and only if the player who sends it subsequently takes an action a in $H(m)$. Hence, H is a correspondence from M to A, which we call *the honesty correspondence,* $H : M \rightrightarrows A$. For any action $a \in A$, let $M_a = \{m \in M \mid a \in H(m)\}$, the set of messages that are honest when a is played.[7]

Suppose that a player has sent a message m and then taken an action $a \in A$. Whether or not a message is honest depends in part on the other player's message, since $a = f(m')$, where m' is the other's message. We say that a player's pure strategy (m, f) is *strictly honest* if $f(m') \in H(m)$ for all $m' \in M$, that is, if the message is honest irrespective of what message the other player sends. Likewise, a mixed strategy $\sigma \in \Delta(S)$ will be called strictly honest if all pure strategies in its support are strictly honest.[8]

Players have a lexicographic *preference for honesty*, defined as follows.[9] First, let $w[(m, f), (m', g)] = 0$ if the player's own strategy is (m, f), the other's is (m', g)), and $f(m') \in H(m)$. Otherwise, $w[(m, f), (m', g)] < 0$. Hence, $-w$ may be thought of as the (psychological) "cost" of dishonesty. With some abuse of notation, let $w(\sigma, \sigma')$ be the *expected* value of w for a player who uses the mixed strategy σ when the other uses σ'. Secondly, let $\succcurlyeq_L$ define the *lexicographic order* on $\mathbb{R}^n$, for any integer $n > 1$, defined as usual, that is, $(x_1, x_2, ...x_n) \succcurlyeq_L (y_1, y_2, ..y_n)$ if $x_1 > y_1$ or $x_1 = y_1$ and

[7] In earlier versions of this paper, we called H the *truth* correspondence, a slightly more restrictive interpretation. The requirement that each message $m \in M$ be either honest (true) or dishonest (untrue) clearly rules out, from the set M, messages such as "This message is dishonest" ("This message is untrue").

[8] People may have different views about what honesty is, in part depending on their culture. While some may think that planning to lie only in case the other player would deviate from the equilibrium path is not dishonest (this is essentially what we assume here), others may require stricter moral standards and deem it dishonest even to plan to lie in situations that occur with probability zero (this is essentially our definition of "strict honesty"). Our results hold for both definitions, see section 5.

[9] For the case of a lexicographic preference for strict honesty, rather than for honesty, see discussion in section 5.

$x_2 > y_2$, etc. Third, define each player's *utility vector,* when the own strategy is σ and the other's is σ', as

$$\tilde{v}(\sigma, \sigma') = (v(\sigma, \sigma'), w(\sigma, \sigma')) \in \mathbb{R}^2. \tag{5}$$

Finally, the ordinal preferences of the players in $\tilde{\mathcal{G}}$ are defined as the lexicographic ordering of these utility vectors. In other words: each player (strictly) prefers one strategy profile over another if the first profile's utility vector is lexicographically ranked (strictly) before the other's:

$$(\sigma, \sigma') \succcurlyeq (\tau, \tau') \quad \Leftrightarrow \quad \tilde{v}(\sigma, \sigma') \succcurlyeq_L \tilde{v}(\tau, \tau'), \tag{6}$$

where σ, $\tau \in \Delta(S)$ are the player's own strategies and σ',$\tau' \in \Delta(S)$ those of the other player. This defines $\tilde{\mathcal{G}} = (S, \succcurlyeq)$ as an *ordinal game.*

We define the best-reply correspondence $\tilde{\beta} : \Delta(S) \rightrightarrows \Delta(S)$ in $\tilde{\mathcal{G}}$ by

$$\tilde{\beta}(\sigma') = \{\sigma \in \Delta(S) : (\sigma, \sigma') \succcurlyeq (\tau, \sigma') \quad \forall \tau \in \Delta(S)\}. \tag{7}$$

A *Nash equilibrium* of the ordinal game $\tilde{\mathcal{G}}$ is a strategy profile (σ, σ') such that $\sigma \in \tilde{\beta}(\sigma')$ and $\sigma' \in \tilde{\beta}(\sigma)$. Such an equilibrium is *symmetric* if $\sigma = \sigma'$. The set of strategies in symmetric Nash equilibria of $\tilde{\mathcal{G}}$ will be denoted

$$\tilde{\Delta}^{NE} = \left\{\sigma \in \Delta(S) : \sigma \in \tilde{\beta}(\sigma)\right\}. \tag{8}$$

The following two axioms turn out to be important and will be explicitly invoked when assumed:

> **Axiom P** (the precision axiom): For each action $a \in A$ there exists at least one message $m \in M$ such that $H(m) = \{a\}$.
>
> **Axiom N** (the null axiom): There exists at least one message $m \in M$ such that $H(m) = A$.

In other words, Axiom P requires the language to be rich enough to contain at least one message for each action of the underlying G such that the action is exactly specified; to send such a message and then take another action is dishonest.[10] Axiom N requires the language to contain messages that do not say anything about what

[10]Likewise, Rabin (1994) defines *completeness* of a pre-play communication language to essentially mean that in the pre-play negotiation stage in his model, players are able to specify any equilibrium than they want to suggest (op. cit. Definition 2).

action in G the speaker will use. Such messages will be called *null messages*. They are always honest, irrespective of the action taken by the player.[11] To send such a message can thus be thought of as sending no message at all.

Remark 1. *We obtain cheap talk as the special case when all messages are null messages: then all messages are always honest.*

3. Nash equilibrium

It follows from the definition of the best-reply correspondence $\tilde{\beta}$ that a mixed-strategy profile (σ, σ) is a Nash equilibrium of $\tilde{\mathcal{G}}$ if and only if (i) it is a Nash equilibrium of $\mathcal{G}$, (ii) all strategies in the support of σ have the same expected cost of dishonesty against σ, and (iii) there is no other pure strategy that earns the same material payoff against σ and has a lower expected cost of dishonesty against σ. Formally (and with a slight abuse of notation):

Lemma 1. *$\sigma \in \tilde{\beta}(\sigma)$ if and only if $\sigma \in \beta(\sigma)$ and*

$$v((m,f),\sigma) = v(\sigma,\sigma) \quad \Rightarrow \quad w((m,f),\sigma) \leq w((m',g),\sigma)$$

for all $(m,f) \in S$ and all $(m',g) \in supp(\sigma)$.

As an immediate corollary we obtain that if (σ, σ) is a Nash equilibrium of $\tilde{\mathcal{G}}$ in which a null message is used with positive probability, then σ has in its support only pure strategies that are honest against σ. We call such equilibria *honesty equilibria*. By contrast, a symmetric Nash equilibrium (σ, σ) of $\tilde{\mathcal{G}}$ is a *dishonesty equilibrium* if σ assigns positive probability to a pair of pure strategies (m, f) and (m', g) such that $f(m') \notin H(m)$. The following example exhibits a dishonesty equilibrium.

Example 1. *Consider the game G defined by the payoff bi-matrix in (1). Let $M = \{$"c", "d"$\}$, where "c" is honest iff c is played, $H($"c"$) = \{c\}$, and "d" is honest iff d is played, $H($"d"$) = \{d\}$. Consider the pure strategy $s = ($"d", $f)$, where $f($"d"$) = c$ and $f($"c"$) = d$. In other words: say "d" and take action c if you receive the message "d", otherwise take action d. Clearly (s, s) is a Nash equilibrium in the cheap-talk game $\mathcal{G}$, since no deviation can result in a higher payoff in G. A deviation to "c" results in a payoff loss in G, so (s, s) is also a Nash equilibrium in the lexicographic communication game $\tilde{\mathcal{G}}$, a dishonesty equilibrium.*

[11] Hurkens and Schlag (2002) analyze cheap talk pre-play communication in situations where each player has the option of not showing up at the pre-play communcation stage, that is, to neither senda message nor know if the other player has sent a message. (By contrast, our players cannot commit not to hear or see the other's message.) They show that the unique evolutionarily stable set in $n \times n$-coordination games is characterized by play of the Pareto dominant equilibrium.

Next, we consider the opposite possibility, discussed in Aumann (1990), namely that people may say "c" even when they intend to play d in the game G in (1). Such behavior, while compatible with Nash equilibrium under cheap talk, is incompatible with Nash equilibrium in **a** lexicographic communication games if saying "c" is dishonest when d is played, and if the language is rich enough to contain a message that is honest when d is played.

Example 2. *Let G be as in the preceding example, let M be any message set and H an honesty correspondence such that $d \in H\left(m^d\right)$ for some message $m^d \in M$. Suppose that (σ, σ) is a dishonest strategy profile resulting in play of (d, d) with probability one, that is, σ assigns positive probability so some message that is dishonest when d is played. A unilateral deviation to the pure strategy $s = \left(m^d, f^d\right)$, where $f^d(m) = d$ for all $m \in M$, incurs no loss of payoff in G, but results in a lexicographic gain due to honesty. Hence, such a profile (σ, σ) is not a Nash equilibrium of the lexicographic communication game $\tilde{\mathcal{G}}$.*

We now explore the implications of Axioms P and N. First, if the language contains a null message, then any symmetric Nash equilibrium of G can be implemented in Nash equilibrium in $\tilde{\mathcal{G}}$ by simply having both players send a null message ("promise nothing") and play the symmetric Nash equilibrium of G irrespective of the message received from the other player. In particular, the Pareto dominated equilibrium (d, d) in the game G in (1) is consistent with Nash equilibrium in $\tilde{\mathcal{G}}$. Denoting mixed strategies in G by μ, with $\mu(a)$ for the probability assigned to action $a \in A$, we have:

Lemma 2. *Let $(\mu, \mu) \in \Delta(A) \times \Delta(A)$ be a Nash equilibrium of G and assume that the language in $\tilde{\mathcal{G}}$ satisfies axiom N. Then there exists a symmetric honesty equilibrium of $\tilde{\mathcal{G}}$ in which each action $a \in A$ is played with probability $\mu(a)$.*

Second, if G is an coordination game with at least two actions then every symmetric Nash equilibrium in the lexicographic communication game has a message that is not sent in equilibrium. More precisely, a finite and symmetric $n \times n$-game G is a (pure) *coordination game* if the payoff matrix Π satisfies $\pi(i, i) > \pi(j, i)\ \forall i, j \neq i$. In other words, each (pure) action then is its own unique best reply. A message $m \in M$ is *unsent* under a mixed strategy $\sigma \in \Delta(S)$ if no pure strategy in the support of σ uses m with positive probability.

Proposition 1. *Let $\tilde{\mathcal{G}}$ be a lexicographic communication game that satisfies Axioms P and N, and where G is an $n \times n-$coordination game with $n \geq 2$. Every $\sigma \in \tilde{\Delta}^{NE}$ has at least one unsent message.*

Proof: Consider a mixed strategy $\sigma \in \Delta(S)$ such that every message $m \in M$ is sent with positive probability in σ. By Axiom N, the language contains a null message. Let m_0 be such a message. Since m_0 is used in σ, no pure strategy (m, f) in the support of σ is dishonest against σ, by Lemma 1. Moreover, since every message is sent with positive probability, the support of σ contains only pure strategies $s = (m, f)$ such that $f(m') \in H(m)$ for all $m' \in M$. By hypothesis, the game G contains at least two actions, say c and d. By Axiom P there exist messages "c","d"$\in M$ such that $H($"c"$) = \{c\}$ and $H($"d"$) = \{d\}$. Under the strategy profile (σ, σ), the pair of messages ("c","d") is realized with positive probability. The player who sent "c" has to play c, but this is not a best reply to the action of the other player, who plays d (since she sent "d"). Hence, $\sigma \notin \tilde{\beta}(\sigma)$. **End of proof.**

The following example shows that there are dishonest equilibria in some games even under the hypotheses of Proposition 1:

Example 3. *Reconsider the game G in (1) and let $M = \{$"c", "d", $m_0\}$, where "c" is honest iff c is played, "d" iff d is played and m_0 is always honest. Consider the pure strategy $s = ($"d", $f)$, where $f($"d"$) = c$ and $f($"c"$) = f(m_0) = d$. In other words: say "d", and take action c if and only if you receive the message "d". Messages "c" and m_0 are thus unsent in s. It is easily verified that (s, s) is a Nash equilibrium of $\tilde{\mathcal{G}}$ for the reasons given in Example 1.*

The next example illustrates that if axiom N is met and there is more than one null message, then Nash equilibrium in the lexicographic communication game permits payoffs that are incompatible with Nash equilibrium in the underlying game G (for reasons given more generally in Banerjee and Weibull (2000)).

Example 4. *Reconsider the game G in (1) and let $M = \{$"c", "d", $m_0, m_1\}$, where "c" is honest iff c is played, "d" iff d is played and m_0 and m_1 are always honest. The expected material payoff 8 is then obtained in symmetric Nash equilibrium of $\tilde{\mathcal{G}}$ as follows. Let σ randomize 50/50 between the two pure strategies $s = (m_0, f)$ and $s^* = (m_1, g)$, where $f(m_1) = c$ and $f(m') = d$ for all $m' \neq m_1$, and $g(m_0) = c$ and $g(m') = d$ for all $m' \neq m_0$. The outcome under (σ, σ) is a randomization over (c, c) and (d, d) with equal probability for each. Hence, the expected payoff in G is 8 and there is no dishonesty. This is a Nash equilibrium in $\tilde{\mathcal{G}}$, because sending messages "c" or "d" only results in a payoff loss in G (meeting action d for sure). Moreover, sending m_0 with probability q and m_1 with probability $1 - q$, for some $q \in [0, 1]$, and*

best-responding to the message sent by σ, results in the following expected payoff in G:

$$7\frac{q}{2} + 9\frac{q}{2} + 9\frac{1-q}{2} + 7\frac{1-q}{2} = 8. \tag{9}$$

Hence, $\sigma \in \tilde{\Delta}^{NE}$.

4. Set-wise evolutionary stability

The concept of *neutral stability* (Maynard Smith (1982)) is a weakening of evolutionary stability: instead of requiring that any mutant strategy does strictly worse in the post-mutation population (granted its population share is small enough) it is required that no mutant does strictly better in the post-mutation population (under the same proviso). Neutral stability is thus similar in spirit to Nash equilibrium in the sense that no small group of individuals in a large community can do better by deviating to another strategy when the rest of the community plays the original strategy.[12] In the context of cheap-talk games, the definition runs as follows:

Definition 1. *$\sigma \in \Delta^{NE}$ is **neutrally stable** in $\mathcal{G}$ if $\forall \tau \in \Delta(S)$:*
(i) $v(\tau, \sigma) < v(\sigma, \sigma)$ or
(ii) $v(\tau, \sigma) = v(\sigma, \sigma) \ \wedge \ v(\tau, \tau) \leq v(\sigma, \tau)$.

Nash equilibria in cheap-talk games are not isolated but form continua. Typically, strategies in the same continuum set of Nash equilibria lead to the same outcome (probability distribution over payoffs). More precisely, the cheap-talk game $\mathcal{G}$ being finite, its non-empty set of mixed-strategy Nash equilibria consists of finitely many disjoint, closed and connected semialgebraic sets, the *Nash equilibrium components* of $\mathcal{G}$.[13] It follows that also the set Δ^{NE} is non-empty and consists of finitely many disjoint, closed and connected semialgebraic subsets, which we will call the *components* of Δ^{NE}.[14]

In the evolutionary paradigm, drift may occur within each component of Δ^{NE}. In order to take account of this possibility, Thomas (1985) suggested a notion of set-wise evolutionary stability, sets of neutrally stable strategies that are robust against drift away from the set. One can show that each minimal evolutionarily stable set

[12] In the case of evolutionary, as opposed to neutral stability, such groups do strictly worse (hence a parallel to strict Nash equilibrium).

[13] See e.g. Kohlberg and Mertens (1986).

[14] The set is a projection of the intersection between the set of Nash equilibria and the diagonal of the space of mixed-strategy profiles. It is non-empty by Kakutani's fixed-point theorem applied to the correspondence β, see Weibull (1995).

coincides with a component of Δ^{NE}.[15] Formally, a set of neutrally stable strategies is called an *evolutionarily stable set* if the weak inequality in condition (ii) in the above definition of neutral stability is strict whenever τ lies outside the set:

Definition 2. *$X \subset \Delta^{NE}$ is an **evolutionarily stable set** in $\mathcal{G}$ if $\forall \sigma \in X, \tau \in \Delta(S)$:*
(i) $v(\tau,\sigma) < v(\sigma,\sigma)$ or
(ii) $v(\tau,\sigma) = v(\sigma,\sigma) \wedge v(\tau,\tau) < v(\sigma,\tau)$ or
(iii) $v(\tau,\sigma) = v(\sigma,\sigma) \ \wedge \ v(\tau,\tau) = v(\sigma,\tau) \wedge \tau \in X$.

4.1. Definitions for lexicographic communication games. In the game-theory literature one finds a variety of distinct but related definitions of neutral stability in lexicographic games, see Rubinstein (1986), Abreu and Rubinstein (1988), Fudenberg and Maskin (1990), Binmore and Samuelson (1992), Rubinstein (2000) and Samuelson and Swinkels (2003). The following definition is close to that in Rubinstein (2000).[16] Before giving the definition, we note, by way of an example, that neutrally stable strategies in cheap-talk games need not even be Nash equilibrium strategies in lexicographic communication games.

Example 5. *Reconsider the lexicographic game $\tilde{\mathcal{G}}$ based on game G in equation (1) and with message set $M = \{$"c", "d", $m_0\}$, where "c" is honest iff c is played, "d" iff d is played, and m_0 is always honest (as in Example 3). As shown in Banerjee and Weibull (2000), the strategy σ that sends all messages with equal probability and replies to the same message with d and to a different message with c is neutrally stable in the cheap-talk game $\mathcal{G}$. However, $\sigma \notin \tilde{\Delta}^{NE}$ since a lexicographically better reply to σ is to send m_0 with probability one, reply to m_0 with d and to the other two messages with c.*

Consider a lexicographic communication game $\tilde{\mathcal{G}}$ defined as above:

Definition 3. *$\sigma \in \tilde{\Delta}^{NE}$ is **neutrally stable** in $\tilde{\mathcal{G}}$ if $\forall \tau \in \Delta(S)$:*
(i) $v(\tau,\sigma) < v(\sigma,\sigma)$ or
(ii) $v(\tau,\sigma) = v(\sigma,\sigma) \ \wedge \ v(\tau,\tau) < v(\sigma,\tau)$ or
(iii) $v(\tau,\sigma) = v(\sigma,\sigma) \ \wedge \ v(\tau,\tau) = v(\sigma,\tau) \wedge \ w(\tau,\sigma) \leq w(\sigma,\sigma)$.

[15] More generally, each evolutionarily stable set is the union of components of Δ^{NE}. See section 2.4 in Weibull (1995).

[16] In comparison with Rubinstein's definition 5 in chapter 2 (op.cit.), we note two differences. First, Rubinstein considers lexicographic costs of complexity while we consider lexicographic costs of dishonesty. Secondly, unlike Rubinstein, we restrict the definition to Nash equilibrium strategies.

Had the cost of dishonesty been zero ($w \equiv 0$), then this definition would have boiled down to that for cheap-talk games, while for positive costs of dishonesty the criterion is more stringent in the lexicographic case. For if the first two equalities in (iii) are met, then τ is required to incur a cost of dishonesty against σ not exceeding that of σ against itself. Hence, neutral stability in $\tilde{\mathcal{G}}$ is a refinement of neutral stability in $\mathcal{G}$. Let $\tilde{\Delta}^{NSS} \subset \tilde{\Delta}^{NE}$ denote the (possibly empty) set of neutrally stable strategies in $\tilde{\mathcal{G}}$.

Neutral stability in $\tilde{\mathcal{G}}$ does not imply Pareto efficiency in Aumann's (1990) example:

Example 6. *The strategy σ in the Example 4 is neutrally stable in $\tilde{\mathcal{G}}$. To see this, recall that $\sigma \in \tilde{\Delta}^{NE}$. By Proposition 3 in Banerjee and Weibull (2000), σ is neutrally stable in the cheap-talk game $\mathcal{G}$. Hence, it remains to show that if (i) and (ii) are not met, then (iii) is satisfied. By the proof of the same proposition: $v(\tau,\sigma) = v(\sigma,\sigma) \wedge v(\tau,\tau) = v(\sigma,\tau)$ implies $\tau = \sigma$. Hence, (iii) is trivially met. We conclude that $\sigma \in \tilde{\Delta}^{NSS}$.*

The following set-wise extension parallels that for cheap-talk games:

Definition 4. *$X \subset \tilde{\Delta}^{NE}$ is an **evolutionarily stable set** in $\tilde{\mathcal{G}}$ if $\forall \sigma \in X, \tau \in \Delta(S)$:*
(i) $v(\tau,\sigma) < v(\sigma,\sigma)$ or
(ii) $v(\tau,\sigma) = v(\sigma,\sigma) \wedge v(\tau,\tau) < v(\sigma,\tau)$ or
(iii) $v(\tau,\sigma) = v(\sigma,\sigma) \ \wedge \ v(\tau,\tau) = v(\sigma,\tau) \wedge \ w(\tau,\sigma) < w(\sigma,\sigma)$ or
(iv) $v(\tau,\sigma) = v(\sigma,\sigma) \ \wedge \ v(\tau,\tau) = v(\sigma,\tau) \wedge \ w(\tau,\sigma) = w(\sigma,\sigma) \wedge \ \tau \in X$.

By definition, each strategy in an evolutionarily stable set in $\tilde{\mathcal{G}}$ is neutrally stable in $\tilde{\mathcal{G}}$.

4.2. Coordination games. Let G be a finite and symmetric coordination game with a unique Pareto dominant Nash equilibrium (c,c).[17] Let α be the payoff to each player in the Pareto dominant Nash equilibrium. Now consider a lexicographic communication game $\tilde{\mathcal{G}}$ based on G. The Pareto optimal outcome in such a game is clearly that both players receive material payoff α and are honest. Let

$$X^* = \{\sigma \in \Delta(S) : v(\sigma,\sigma) = \alpha \text{ and } w(\sigma,\sigma) = 0\}. \tag{10}$$

[17] That is, (c,c) is a Nash equilibrium of G and both players obtain lower payoffs in all other Nash equilibria of G.

In general, this set is not a singleton. It may include messages expressing precise, vague or no intentions as to actions in the underlying game, as well as strategies that vary in their response to some messages.

By Pareto dominance, each $\sigma \in X^*$ is a best reply to itself, so $X^* \subset \tilde{\Delta}^{NE}$. We proceed to establish that if the language in the communication game $\tilde{\mathcal{G}}$ satisfies Axioms P and N, then X^* is an evolutionarily stable set in $\tilde{\mathcal{G}}$ and all other sets $X \subset \tilde{\Delta}^{NE}$ are evolutionarily unstable in $\tilde{\mathcal{G}}$, granted the underlying game contains at least two actions. Formally:

Proposition 2. *Let G be a finite and symmetric coordination game with at least two actions and with a unique Pareto dominant Nash equilibrium. If $\tilde{\mathcal{G}}$ is a lexicographic communication game based on G, satisfying Axioms P and N, then X^* is the unique evolutionarily stable set in $\tilde{\mathcal{G}}$.*

Proof: We prove first that X^* is an evolutionary stable set. Suppose, thus, that $\sigma \in X^*$, and let $\tau \in \Delta(S)$. We proceed to show that one of conditions (i)-(iv) in definition 4 is met. Since $X^* \subset \tilde{\Delta}^{NE}$: $v(\tau, \sigma) \leq v(\sigma, \sigma)$. If the inequality is strict, condition (i) is met. In case of equality, we have $v(\tau, \sigma) = v(\sigma, \sigma) = \alpha$. But, since α is the Pareto efficient payoff, τ must send only messages that make σ react with c, and τ must always reply to messages from σ by taking action c. Hence, $v(\sigma, \tau) = \alpha$, and thus $v(\tau, \tau) \leq v(\sigma, \tau)$ by Pareto dominance. If this inequality is strict, (ii) is met. If equality holds, then $v(\tau, \tau) = \alpha$, which implies, as before, that τ reacts with c to itself. Now σ is honest against itself, so $w(\tau, \sigma) \leq w(\sigma, \sigma)$. If this inequality is strict, then (iii) is met. In case of equality, $w(\tau, \sigma) = w(\sigma, \sigma)$ implies that τ is honest against σ. However, when τ meets σ, it takes action c with probability one, so τ satisfies $M(\tau) \subset M_c$. But since τ always plays c against itself, this means that it is honest against itself; $w(\tau, \tau) = 0$. Hence, $\tau \in X^*$, and thus condition (iv) is met.

Next, we prove that X^* is the only evolutionarily stable set. Hence, suppose that $X \subset \tilde{\Delta}^{NE}$ is evolutionarily stable and, contrary to the claim, there exists a strategy $\sigma \in X$ such that $v(\sigma, \sigma) < \alpha$ or $w(\sigma, \sigma) < 0$ (or both). Suppose that $v(\sigma, \sigma) < \alpha$. Since $\sigma \in \tilde{\Delta}^{NE}$ and axioms N and P hold, proposition 1 implies that there exists a message, say m^0, that is not sent in σ. We proceed in three steps to show that this leads to a contradiction. First, we construct a strategy σ^0 that does not send m^0, behaves like σ against strategies in X, and is "nice" to senders of m^0. Secondly, we show that $\sigma^0 \in X$. Thirdly, we show that $\sigma^0 \notin \tilde{\Delta}^{NE}$, contradicting the hypothesis $X \subset \tilde{\Delta}^{NE}$.

Step 1: For any pure strategy $s = (m, f) \in supp(\sigma)$, let s^0 be the associated modified pure strategy (m, g), where $g(m) = f(m)$ for all $m \neq m^0$ and $g(m^0) = c$. If $\sigma = \sum \lambda_i s_i$ for some pure strategies s_i, and probability weights $\lambda_i > 0$ summing to 1, denote by σ^0 the associated sum $\sum \lambda_i s_i^0$. In other words, σ^0 is the same convex combination of the pure strategies s_i^0 as σ is with respect to the s_i. Message m^0 is never sent by σ nor by σ^0, and σ and σ^0 differ only in their reaction to m^0, so their payoffs when playing against each other are the same:

$$v(\sigma^0, \sigma) = v(\sigma, \sigma) = v(\sigma^0, \sigma^0) = v(\sigma, \sigma^0) \tag{11}$$

and

$$w(\sigma^0, \sigma) = w(\sigma, \sigma) = w(\sigma^0, \sigma^0) = w(\sigma, \sigma^0). \tag{12}$$

Step 2: By condition (iv) in definition 4, $\sigma^0 \in X$.

Step 3: $\sigma^0 \notin \tilde{\Delta}^{NE}$ since the pure strategy $r = (m^0, f^0)$, where $f^0(m) \equiv c$ is a better reply to σ^0: $v(\sigma^0, \sigma^0) < \alpha = v(r, \sigma^0)$.

Finally, suppose that $v(\sigma, \sigma) = \alpha$ and $w(\sigma, \sigma) < 0$. By Lemma 1 and its corollary, no null message is sent in σ. Now the proof runs as before: let σ^0 be defined as in step 1 above, this time applied to m^0, when this is a null message. A repetition of the arguments in steps 1 and 2 shows that $\sigma^0 \in X$. Moreover σ^0 plays c against message m^0. But then $\sigma^0 \notin \tilde{\Delta}^{NE}$ since the pure strategy r defined in step 3 is a lexicographically better reply to σ^0: $v(r, \sigma^0) = v(\sigma^0, \sigma^0)$ and $w(m^0, \sigma^0) = 0 > w(\sigma, \sigma) = w(\sigma^0, \sigma^0)$, contradicting the hypothesis $X \subset \tilde{\Delta}^{NE}$. **End of proof.**

Remark 2. *It is easily verified that the same conclusion holds under the weaker hypothesis that the base game G need not be a coordination game, but has a symmetric and strict Nash equilibrium that Pareto dominates all other outcomes in G.*

Remark 3. *We want to point out a form of non-robustness of the claim in Proposition 2 with respect to players' preferences. Suppose that, instead of a lexicographic preference for honesty, players have additively separable payoffs of the form $u(\sigma, \sigma') = v(\sigma, \sigma') + \varepsilon w(\sigma, \sigma')$ for $\varepsilon > 0$, where $w(\sigma, \sigma')$ is the expected value of the function w defined by $w[(m, f), (m', g)] \in \{0, -1\}$ with $w[(m, f), (m', g)] = 0$ if and only if $f(m') \in H(m)$ (here (m, f) is the player's own strategy while (m', g) is that of the other player). For small ε, such non-cheap-talk games can be thought of as approximations of lexicographic communication games. By way of proposition 3 in Banerjee and Weibull (2000) and a straight-forward perturbation argument, it is not difficult to show that for symmetric 2×2-coordination game with strict equilibrium*

payoffs α and $\beta < \alpha$, and for all $\varepsilon > 0$ sufficiently small, there exists exactly one more evolutionarily stable set. Moreover, each strategy σ in that set earns an expected payoff close to $\alpha - (\alpha - \beta)/|M|$ against itself (this is the limit payoff as $\varepsilon \to 0$). So for small material costs for dishonesty, unlike lexicographic costs, communication does not lead to full efficiency.[18] *However, this non-robustness is insignificant in terms of outcomes when the message set M is large, an arguably relevant case for applications to communication in natural language.*

5. Concluding comments

While we defined honesty as a property of a message in relation to actual behavior, a taken action, strict honesty was defined as a property of a message in relation to an "intention," a contingent plan for the action to take in different potential circumstances (here, for all messages sent by the other player). More precisely, in section 2 we defined a pure strategy (m, f) to be *strictly honest* if $f(m') \in H(m)$ for all $m' \in M$, and a mixed strategy $\sigma \in \Delta(S)$ to be strictly honest if all pure strategies in its support are strictly honest. In other words, while a preferences for honesty can be thought of as a desire to avoid (own or others') disapproval of one's actions ("shame"), a preference for strict honesty can be thought of as a desire to avoid (own) disapproval of one's plans or intentions ("guilt"). Indeed, existing moral codes and religions, such as Catholicism and Protestantism, arguably take distinct position on whether acts or intentions count.

Suppose now that players, rather than having a lexicographic preference for honesty, have a lexicographic preference for strict honesty, defined as follows. First, let $w(m, f) = 0$ if the player's strategy is (m, f) is strictly honest, otherwise $w(m, f) < 0$. Hence, w is now the cost of not being strictly honest. Proceeding as in the case of honesty, this results in a lexicographic communication game, $\bar{\mathcal{G}}$, that is easier to analyze than $\tilde{\mathcal{G}}$, since strict honesty is a property of a strategy while lying is a property of a strategy in the context of a strategy pair . Indeed, it is not difficult to verify that Proposition 2 is valid also for $\bar{\mathcal{G}}$. Roughly speaking, a lexicographic preference for strict honesty provides a more direct and slightly more stringent selection against dishonesty.

An interesting feature of evolutionary stability in pre-play communication games is that ordinality—that is, invariance of the solution set under transformations that leave the best-reply correspondence unchanged—turns out not to be robust to the

[18] See Binmore and Samuelson (1992), Rubinstein (2000) and Samuelson and Swinkels (2003) for similar robustness analyses.

introduction of a pre-play communication stage. For example, while the best-reply correspondence of the game in (1) is identical with that in

$$\begin{array}{cc} c & d \\ c \;\; 1,1 & 0,0 \\ d \;\; 0,0 & 7,7 \end{array} \tag{13}$$

the unique evolutionarily stable set in a lexicographic pre-play communication game, as modelled above, results in play of (c, c) when based on (1) but on (d, d) when based on (13).[19] When switching from (1) to (13), the best-reply correspondence of the lexicographic communication game changes, and hence ordinality does not require that the solution be unaffected. A more profound question is whether ordinality should be viewed as a general desideratum for solution concepts — a question that falls outside the scope of this study.

When applying our model to Aumann's example, we came to the conclusion that the outcome (c, c), which Aumann convincingly argues is not self-enforcing when players have no preference for honesty *per se*, is the only robust outcome "in the long run." Expressed somewhat loosely: if such a game were played with pre-play communication, over and over again in a large population with a common language and where people have a lexicographic preference for honesty, then agreement to play (c, c) and to honor that agreement would be the only mode of behavior that would be sustainable in the long run. By the same token: while the Nash equilibrium (d, d) is arguably self-enforcing in the absence of a preference for honesty, in the sense of Aumann (1990) (though no such claim is explicitly made by Aumann), it is not a robust long-run outcome, according to our analysis. When players have a lexicographic preference for honesty, such a population, even if it were initially playing (d, d), would eventually find its way to the Pareto efficient outcome (c, c). Our theoretical results are in good agreement with the empirical finding in Blume and Ortman (2005). Based on laboratory experiments, the authors conclude that, in games with payoff structures similar to that in Aumann's example, costless communication with *a priori* meaningful messages leads to the efficient outcome after some rounds of play. In a follow-up on Gneezy (2005), Hurkens and Kartik (2006) find that Gneezy's data cannot reject the hypothesis that some people never lie while others lie whenever they obtain a material benefit from that. In particular, an individual's propensity to lie may not depend on the individual's material benefit nor on the harm done to others.

[19] Note, however, that evolutionary and neutral stability are ordinal solution concepts in the sense of being invariant under transformations that keep the best-reply correspondence unchanged.

To us, this seems to lend some empirical support to the here maintained hypothesis of a (perhaps culturally conditioned) lexicographic preference against lying, as opposed a trade-off between honesty and material payoffs.

We plan to extend our analysis in two directions. One extension is to study the implications of the present approach for infinitely repeated games (cf. Fudenberg and Maskin (1990)). The second is to enrich the language and the correspondence H to allow for conditional statements, that is, statements the honesty (or truth) of which may depend on the other player's message. This extension, though potentially difficult, appears particularly relevant for games with asymmetric equilibria that are Pareto efficient, allowing players to correlate their play across such equilibria.

References

[1] Abreu D. and A. Rubinstein (1988): "The structure of Nash equilibrium in repeated games with finite automata", *Econometrica* 56, 1259-1282.

[2] Alger I. and A. Ma (2003): "Moral hazard, insurance and some collusion", *Journal of Economic Behavior and Organization* 50, 225-247.

[3] Alger I. and R. Renault (2006a): "Screening ethics when honest agents care about fairness", *International Economic Review* 47, 59-85.

[4] Alger I. and R. Renault (2006b): "Screening ethics when honest agents keep their word", forthcoming, *Economic Theory*.

[5] Aumann R. (1990): "Nash equilibria are not self-enforcing", chapter 34 in J. Gabszewicz, J.-F. Richard and L. Wolsey, *Economic Decision Making: Games, Econometrics, and Optimization*. Elsevier Science Publishers.

[6] Banerjee A. and J. Weibull (2000): "Neutrally stable outcomes in cheap-talk coordination games", *Games and Economic Behavior* 32, 1-24.

[7] Benabou R. and G. Laroque (1992): "Using priviliged information to manipulate markets: Insiders, gurus and credibility", *Quartely Journal of Economics* 107, 921-958.

[8] Benoit J.-P. and V. Krishna (1993): "Renegotiation in finitely repeagted games", *Econometrica* 61, 303-323.

[9] Binmore K. and L. Samuelson (1992): "Evolutionary stability in repeated games played by finite automata", *Journal of Economic Theory* 57, 278-305.

[10] Binmore K. and L. Samuelson (1994): "Drift", *European Economic Review* 38, 859-867.

[11] Binmore K. and L. Samuelson (1997): "Muddling through: Noisy equilibrium selection", *Journal of Economic Theory* 74, 235-265.

[12] Blume A. (1998): "Communication, risk, and efficiency in games", *Games and Economic Behavior* 22, 171-202.

[13] Blume A., Y.-G. Kim and J. Sobel (1993): "Evolutionary Stability in games of communication", *Games and Economic Behavior* 5, 547-575.

[14] Blume A. and A. Ortman (2005): "The effect of costless pre-play communication: experimental evidence for games with Pareto-ranked equilibria", forthcoming, *Journal of Economic Theory.*

[15] Charness G. (2000): "Self-serving cheap talk: a test of Aumann's conjecture", *Games and Economic Behavior* 33, 177-194.

[16] Clark K., S. Kay and M. Sefton (2001): "When are Nash equilibria self-enforcing? An experimental analysis", *International Journal of Game Theory* 29, 495-515.

[17] Conlisk J. (2001): "Costly predation and the distribution of competence", *American Economic Review* 91, 475-484.

[18] Cooper R., D. deJong, R. Forsythe and T. Ross (1989): "Communication in the battle of the sexes game; some experimental results", *RAND Journal of Economics* 20, 568-587.

[19] Crawford V. (1998): "A survey of experiments on communication via cheap talk", *Journal of Economic Theory* 78, 286-298.

[20] Crawford V. (2003): "Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions", *American Economic Review* 93, 133-149.

[21] Ellingsen T. and M. Johannesson (2004): "Promises, threats and fairness", *Economic Journal* 114, 397-420.

[22] Farrell J. (1988): "Communication, coordination, and Nash equilibrium", *Economics Letters* 27, 209-214.

[23] Farrell J. (1993): "Meaning and credibility in cheap-talk games", *Games and Economic Behavior* 5, 514-531.

[24] Farrell J. and R. Gibbons (1989): "Cheap talk with two audiences", *American Economic Review* 79, 1214-1223.

[25] Fudenberg D. and E. Maskin (1990): "Evolution and cooperation in noisy repeated games", *American Economic Review, Papers and Proceedings* 80, 274-279.

[26] Gneeze U. (2005): "Deception: The role of consequences", *American Economic Review* 95, 384-394.

[27] Hurkens S. and K. Schlag (2002): "Evolutionary insights on the willingness to communicate", *International Journal of Game Theory* 31, 511-526.

[28] Hurkens S. and N. Kartik (2006): "(When) Would I lie to you? Comment on 'Deception: the role of consequences'", mimeo., Institut d'Analisi Economica and UCSD.

[29] van Huyck J., R. Battalio and R. Beil (1990): "Tacit coordination games, strategic uncertainty and coordination failure", *American Economic Review* 80, 234-248.

[30] Kohlberg E. and J.-F. Mertens (1986): "On the strategic stability of equilibria", *Econometrica* 54, 1003-1037.

[31] Miettinen T. (2006): "Promises and conventions - an approach to pre-play agreements", mimeo., University College London.

[32] Myerson R. (1989): Credible negotiation statements and coherent plans", *Journal of Economic Theory* 48, 264-303.

[33] Nash J. (1950): "Non-cooperative games", Ph D thesis, Department of Mathematics, Princeton University.

[34] Osborne M. and A. Rubinstein (1994): *A Course in Game Theory*, MIT Press.

[35] Rabin M. (1994): "A model of pre-play communication", *Journal of Economic Theory* 63, 370-391.

[36] Robson A. (1990): Efficiency in evolutionary games: Darwin, Nash and the secret handshake", *Journal of Theoretical Biology* 144, 379-396.

[37] Rubinstein A. (1986): "Finite automata play the repeated prisoners' dilemma", *Journal of Economic Theory* 39, 83-96.

[38] Rubinstein A. (2000): *Economics and Language*. Cambridge University Press.

[39] Samuelson L. and J. Swinkels (2003): "Evolutionary stability and lexicographic preferences", *Games and Economic Behavior* 44, 332-342.

[40] Sobel J. (1985): "A theory of credibility", *Review of Economic Studies* 52, 557-573.

[41] Thomas B. (1985): "On evolutionarily stable sets", *Journal of Mathematical Biology* 22, 105-115.

[42] Weibull J. (1995): *Evolutionary Game Theory*. MIT Press.