Language, meaning and games A model of communication, coordination and evolution

Stefano Demichelis, Jörgen Weibull

ECOLE POLYTECHNIQUE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

LANGUAGE, MEANING AND GAMES

A model of communication, coordination and evolution

Stefano DEMICHELIS
Jörgen W. WEIBULL

January 2008
(Revised version)

Cahier n° 2008-25

DEPARTEMENT D'ECONOMIE
Route de Saclay
91128 PALAISEAU CEDEX
(33) 1 69333033
http://www.enseignement.polytechnique.fr/economie/
mailto:chantal.poujouly@polytechnique.edu

LANGUAGE, MEANING AND GAMES
A model of communication, coordination and evolution

Stefano DEMICHELIS[1]
Jörgen W. WEIBULL[2]

January 2008
(Revised version)

Cahier n° 2008-25

Abstract: Language is arguably a powerful coordination device in real-life interactions. We here develop a game-theoretic model of pre-play communication that generalizes the cheap-talk approach by way of introducing a meaning correspondence between messages and actions, and postulating two axioms met by natural languages. Deviations from this correspondence are called dishonest and players have a lexicographic preference for honesty, second to material payoffs. The model is first applied to two-sided preplay communication in finite and symmetric two-player games and we establish that, in generic and symmetric n × n - coordination games, a Nash equilibrium component in such a lexicographic communication game is evolutionarily stable if and only if it results in the unique Pareto efficient outcome of the underlying game.
We extend the approach to one-sided communication in finite, not necessarily symmetric, two-player games.

JEL code C72, C73, D01.
.

[1] Department of Mathematics, University of Pavia, Italy
[2] Department of Economics, Stockholm School of Economics, Deparment of Economics, Ecole Polytechnique, Paris

LANGUAGE, MEANING AND GAMES

- A model of communication, coordination and evolution

Stefano Demichelis* and Jörgen W. Weibull†

First draft February 1, 2006. This version: January 8, 2008.

Abstract. Language is arguably a powerful coordination device in real-life interactions. We here develop a game-theoretic model of pre-play communication that generalizes the cheap-talk approach by way of introducing a meaning correspondence between messages and actions, and postulating two axioms met by natural languages. Deviations from this correspondence are called dishonest and players have a lexicographic preference for honesty, second to material payoffs. The model is first applied to two-sided preplay communication in finite and symmetric two-player games and we establish that, in generic and symmetric $n \times n$ -coordination games, a Nash equilibrium component in such a lexicographic communication game is evolutionarily stable if and only if it results in the unique Pareto efficient outcome of the underlying game. We extend the approach to one-sided communication in finite, not necessarily symmetric, two-player games.

JEL-codes: C72, C73, D01.

1. Introduction

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

*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.

†Department of Economics, Stockholm School of Economics. Both authors thank Cedric Argenton, Robert Aumann, Milo Bianchi, Vince Crawford, Tore Ellingsen, Ernst Fehr, Drew Fudenberg, Segismundo Izquierdo, Michael Kosfeld and Robert Östling for comments.

[1]Indeed, laboratory experiments usually support the hypothesis that pre-play communication enhances coordination on payoff domant 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.

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

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

This two-player game has three Nash equilibria, all symmetric: the payoff dominant but risk dominated strict equilibrium (c, c), the risk dominant but payoff 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 payoff dominant Nash equilibrium (c, c) is not self-enforcing.

This line of reasoning abstracts away from the possibility that Alice and Bob may have a preference against dishonesty (here, for violating an agreement). In this abstraction, Aumann is not alone. Indeed, virtually all of economics relies on the presumption that economic agents have no preference for honesty or against deceiving or lying *per se*. The standard assumption is that economic agents opportunistically misreport their private information whenever they believe it is to their advantage to do so.[2]

We here show that "small lying costs," in the sense of a lexicographic preference for honesty—when it doesn't reduce material payoffs—render the "bad" equilibrium (d, d) in the above game evolutionarily unstable under two-sided pre-play communication. While small lying costs don't eliminate all bad equilibria, they do destabilize

[2]Notable exceptions are Alger and Ma (2003), Alger and Renault (2006), Alger and Renault (2007), and Kartik, Ottaviani and Squintani (2007).

payoff dominated equilibrium outcomes, where stability is defined in a standard evolutionary model with a set-valued notion of evolutionary stability. When applying our model to Aumann's example, we come to the conclusion that the outcome (c, c), which Aumann convincingly argues is not self-enforcing when players are indifferent towards honesty, is the only robust long-run outcome. 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 a lexicographic preference for honesty, then play of (c, c) would be the only mode of behavior that would be sustainable in the long run. Even if the population were initially playing (d, d), it would eventually find its way to the payoff dominant equilibrium (c, c).

More precisely, we generalize the cheap-talk approach to include what we call a *meaning correspondence*, a correspondence that specifies what pre-play messages mean in terms of the action to be taken in the underlying game, such as the one in (1). For instance, the message "I will play c" would typically mean that the sender intends to take action c. To take any other action would be deemed dishonest. By contrast, the message "I will play c or d" is consistent with any action in the game (1) and is thus honest irrespective of what action the sender takes.[3] The key assumption is here that the two parties have a common language and agree on its meaning. Our analysis shows how such a shared culture—language and honesty code—facilitates coordination on socially efficient equilibrium outcomes in strategic interactions. It does not imply honesty, however. Individuals may lie in equilibrium, even when this is part of an evolutionarily stable set. It is rather the common understanding of the language—the common meaning correspondence—that drives home the result.

Most individuals arguably feel some guilt or shame when lying or being dishonest. The practice of using the polygraph in trials suggests that lying causes physiological symptoms of effort (sweating) and recent fMRI studies provide neurological evidence that lying activates more parts of the brain, and parts more associated with negative emotions, than truth-telling.[4] Gneezy (2005) provides experimental evidence for a psychological cost associated with the act of lying, see also Ellingsen and Johannesson (2004), Hurkens and Kartik (2006) and Lundquist *et al.* (2007)). Gneezy's main empirical finding is that "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

[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.

[4]Kozel, Padgett and George (2004) find that "For lying, compared with telling the truth, there is more activation in the right anterior cingulate, right inferior frontal, right orbitofrontal, right middle frontal, and left middle temporal areas." (op.cit., p 855). Other studies suggest that activities in the right side of the brain are correlated with negative emotions, see e.g. Davidson and Hugdahl (1995).

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 in comparison to the material stakes?

This is exactly what we analyze here. We go to the extreme and assume that players avoid dishonest messages only if this comes at *no* loss of material payoff. 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 say that they will play c, but take action d. Such behavior is compatible with Nash equilibrium under cheap talk, since then messages have no exogenous meaning. By contrast, it is incompatible with Nash equilibrium in a lexicographic communication game if the message space is rich enough to permit precise descriptions of actions in the game. 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 is lexicographically better to say m' instead of m, since this can induce no payoff loss in the game G in (1).[5]

Lexicographic preferences for honesty, by themselves, imply neither honesty nor efficiency in equilibrium. In fact, we show that there are Nash equilibria in lexicographic communication games in which both players are dishonest and we also show that there are Nash equilibria in such games that result in outcomes that are payoff dominated by other Nash equilibria in the underlying game G. However, Nash equilibria in pre-play communication games usually come in whole continuum sets, so-called equilibrium components. Our main result is that in finite and symmetric two-player $n \times n$-coordination games with a unique payoff-dominant equilibrium, components that yield payoff dominated outcomes are set-wise evolutionarily unstable, granted the message space satisfies two axioms—a precision and a null axiom—that are met by natural languages. The precision axiom requires that there for each action in the underlying game G exists a message that means that the sender intends to take precisely that action. The null axiom requires that there is a message that means that the sender may take any action. We also show that the payoff-dominant Nash equilibrium outcome is evolutionarily stable. We extend our model to sender-receiver games and show that the sender's most preferred equilibrium is selected. This finding is in agreement with earlier results based on different approaches from ours.

The mechanism that drives home our inefficiency result—that inefficiency leads to evolutionary instability—is similar to that in Robson (1990) in that it depends on the existence of unsent messages in equilibrium. Robson noted that, in a population playing such an equilibrium, a small group of deviating players can profitably use such

[5] 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.

messages as a "secret handshake" to recognize each other and to coordinate their play to an efficient equilibrium. However, while the existence of such unsent messages is assumed in Robson (1990), and non-deviating players in his setting are assumed not to react to these, the existence of unsent messages is here derived from primitives and non-deviators may recognize and even "punish" senders of such messages.

We believe that setwise evolutionary stability is relevant in the present context. If an interaction takes place over and over again in a large population with a common language and culture, then drift may occur among materially payoff-equivalent strategies in connected sets.[6] Thus, if the set is evolutionarily unstable, a small group of individuals can, sooner or later, deviate to some strategy outside the set and do better in terms of material payoffs.

Our results appear to be broadly in agreement with recent empirical findings. Based on laboratory experiments, Blume and Ortman (2005) find 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. To us, this seems to lend some empirical support to the here maintained hypothesis of a (probably culturally conditioned) lexicographic deontological preference for honesty.

The rest of the paper is organized as follows. The model is laid out in section 2, Nash equilibrium is analyzed in section 3 and evolutionary stability in section 4. Section 5 analyzes one-sided communication, section 6 discusses related research and section 7 concludes. Mathematical proofs are given in an appendix.

2. Lexicographic communication games

Let G be a symmetric $n \times n$ two-player game with payoff matrix $\Pi = (\pi(a, b))$. Thus, $\pi(a, b)$ is the payoff to a player who uses pure strategy a when the other player uses pure strategy b. 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*. There is no restriction on what these messages are, but we take them to be statements in a natural language (allowing for basic notation from mathematics), mastered by both persons playing the game in question, and referring to actions to be taken in the game G. The messages can be unconditional, such as "I will take action $a \in A$", or conditional, such as "I will take action $a \in A$ if you say that you

[6] Drift in equilibrium components of games is analyzed in detail in Binmore and Samuelson (1994, 1997), see also Gilboa and Matsui (1991) for the related concept of cyclically stable sets.

will take action a".[7]

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 from the set M to each other. Then each player observes both messages and takes an action $a \in A$. The pure-strategy set for each player in $\mathcal{G}$ is thus the finite set S of pairs (m, f), where $m \in M$ is a message to send and $f : M^2 \longrightarrow A$ a function or "rule" that specifies what action $a = f(m, m')$ in game G to take after having sent message m and received message m', for all possible message pairs (m, m').[8] Given a mixed strategy $\sigma \in \Delta(S)$, a randomization over one's set S of pure strategies, let $\sigma(m, f)$ denote the probability assigned to the pure strategy $s = (m, f)$.[9] Define the payoff function $v : S^2 \to \mathbb{R}$, in $\mathcal{G} = (S, v)$, by letting $v[(m, f), (m', g)]$ be the payoff $\pi[f(m, m'), g(m', m)]$ that a player who takes the action (m, f) against action (m', g) in the underlying game G. We extend the pure-strategy payoff function v linearly to mixed strategies in $\mathcal{G}$ as usual.[10]

Having defined the cheap-talk game $\mathcal{G} = (S, v)$, let $\beta : \Delta(S) \rightrightarrows \Delta(S)$ be the best-reply correspondence in $\mathcal{G}$. This correspondence specifies, for each (pure or mixed) strategy $\sigma' \in \Delta(S)$ that one's opponent may play, the (non-empty) set $\beta(\sigma') \subset \Delta(S)$ of optimal (pure and mixed) strategies to use. Let

$$\Delta^{NE} = \{\sigma \in \Delta(S) : \sigma \in \beta(\sigma)\} \tag{2}$$

be the set of fixed points under β; the set of (pure and mixed) strategies in the cheap-talk game that are best replies to themselves. In other words, Δ^{NE} is the set of pure and mixed strategies used in symmetric Nash equilibria in $\mathcal{G}$.

We are now in a position to define *lexicographic communication games*. The messages, actions and strategies in such a game $\tilde{\mathcal{G}}$ are defined as in $\mathcal{G}$, with G denoting the underlying game. 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)). Messages in $\tilde{\mathcal{G}}$ have a

[7] Note that it is not clear what actions two persons will take who send this conditional statement. However, this would have been clear had they both sent the following message: "I will take action a if also you send this message".

[8] It is technically inessential that each player conditions his action upon his own message (he knows what message he has sent). However, this formalization simplifies the notation.

[9] Technically, $\Delta(S)$ is thus the unit simplex of probability distributions over S. Recall that mixed strategies have two distinct interpretations in game theory. In the *epistemic* interpretation (Aumann and Barndenburger, (1995)), a mixed strategy represents *another* players' uncertainty about the player's behavior . In the *mass action* interpretation (Nash, 1950), there is a population associated with each player role in the game, and a mixed strategy represents a population frequency of deterministic behaviors.

[10] This is done as follows: multiply each pure-strategy payoff $v[(m, f), (m', g)]$ by the probabilities $\sigma(m, f)$ and $\sigma'(m', g)$ attached to the pure strategies involved, and take the sum all these products.

pre-determined meaning in the sense that to send any message $m \in M$ "means" that one intends to take *some* action in a subset of A that depends on m and that may also depend on the message m' received. Let this subset be denoted $\mu(m.m')$. For example, to send the message $m^c =$"I will take action c" would usually be taken to mean that the sender intends to take action c, irrespective of the message received: $\mu(m^c, m') = \{c\}$ for all $m' \in M$. Likewise, the meaning of the message $m^{cd} =$"I will take action c or d" can be formalized as $\mu\left(m^{cd}, m'\right) = \{c, d\}$ for all $m' \in M$. If m^* is the conditional statement "I will take action c if you say that you will take action c" satisfies $\mu(m^*, m^c) = \{c\}$ and $\mu\left(m^*, m^d\right) = A$ for $m^d =$"I will take action d" for any action $d \neq c$.[11] We call such a correspondence $\mu : M^2 \rightrightarrows A$, mapping message pairs to subsets of actions, a *meaning correspondence.*[12]

Players have a lexicographic preference for honesty, defined as follows. Let $h : M^2 \times A \to \mathbb{R}_+$ be the "honesty cost" (psychological and/or social discomfort) of sending message m and taking action a, having received message m', where $h(m, m', a) = 0$ if and only if $a \in \mu(m, m')$, that is, to take actions in accordance with the common language has zero honesty cost, while all other behaviors have positive honesty cost.[13] Define the *second-order payoff* function $w : S^2 \to \mathbb{R}$ by setting $w[(m, f), (m', g)] = -h[m, m', f(m, m')]$. The function value $w[(m, f), (m', g)] \leq 0$ is the second-order utility arising from potentially being dishonest when using pure strategy $(m, f) \in S$ when the other player uses pure strategy $(m', g) \in S$, as, for example, when first saying that one will take a certain action a and then not doing so. With some abuse of notation, let $w(\sigma, \sigma')$ be the linear extension of w to mixed strategies, hence, representing the expected value of w for a player who uses the mixed strategy σ when the other uses σ'. Let $\succcurlyeq_L$ define the *lexicographic order* on $\mathbb{R}^2$, defined as usual: $(x_1, x_2) \succcurlyeq_L (y_1, y_2)$ if $x_1 > y_1$ or $x_1 = y_1$ and $x_2 \geq y_2$. Each player's *utility vector,* when the own strategy is σ and the other's is σ', is defined as

$$\tilde{v}(\sigma, \sigma') = (v(\sigma, \sigma'), w(\sigma, \sigma')) \in \mathbb{R}^2. \tag{3}$$

The preferences of the players in $\tilde{\mathcal{G}}$ are defined as the lexicographic ordering of these utility vectors. In other words: each player prefers one strategy profile over another

[11] Although this does not follow from predicate logic, we conjecture that a vast majority of English-speking persons would understand m^* to also satisfy $\mu(m^*, m^*) = \{c\}$ (or at least $c \in \mu(m^*, m^*)$).

[12] Usually, correspondences are taken to be non-empty valued. However, since there are statements that are dishonest irrespective of the actions taken (for example: "I am a violinist" if uttered by any one of the authors), we allow for the possibility that $\mu(m, m') = \varnothing$ for some $m, m' \in M$. However, by requiring all messages in the set M to be either honest or dishonest, we exclude from the set M such messages as "This message is dishonest", which, arguably, is neither honest nor dishonest.

[13] Individuals may differ as to their honesty costs. The key assumption is that they have a common meaning correspondence.

if the first profile's utility vector is lexicographically ranked before the other's,

$$(\sigma, \sigma') \succcurlyeq (\tau, \tau') \quad \Leftrightarrow \quad \tilde{v}(\sigma, \sigma') \succcurlyeq_L \tilde{v}(\tau, \tau'), \tag{4}$$

where σ, $\tau \in \Delta(S)$ are the player's own strategies and σ',$\tau' \in \Delta(S)$ those of the other player. Material payoffs are thus ranked first and honesty payoffs second. One strategy profile is thus strictly preferred over another if and only if either (i) the expected payoff from the interaction in the underlying game G is higher under the first profile, or (ii) there is an exact tie between those expected payoffs but the expected dishonesty cost is lower under the first profile. This defines $\tilde{\mathcal{G}} = (S, \succcurlyeq)$ as an *ordinal game*.

The best-reply correspondence $\tilde{\beta} : \Delta(S) \rightrightarrows \Delta(S)$ in a lexicographic communication $\tilde{\mathcal{G}}$ is defined by

$$\tilde{\beta}(\sigma') = \{\sigma \in \Delta(S) : (\sigma, \sigma') \succcurlyeq (\tau, \sigma') \quad \forall \tau \in \Delta(S)\}. \tag{5}$$

In other words, a (pure or) mixed strategy σ is a best reply in $\tilde{\mathcal{G}}$ against the pure or mixed strategy σ' if and only if there is no other pure or mixed strategy τ that either results in a higher expected material payoff or in exactly the same material payoff but a lower expected honesty cost. Accordingly, a *Nash equilibrium* of $\tilde{\mathcal{G}}$ 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 used in symmetric Nash equilibria of $\tilde{\mathcal{G}}$ will be denoted

$$\tilde{\Delta}^{NE} = \left\{\sigma \in \Delta(S) : \sigma \in \tilde{\beta}(\sigma)\right\}. \tag{6}$$

This is the set of (pure and) mixed strategies that are best replies to themselves in the lexicographic communication game.

The following two axioms for the meaning correspondence 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 $\mu(m, m') = \{a\}$ for all $m' \in M$.
>
> **Axiom N** (the null axiom): There exists at least one message $m \in M$ such that $\mu(m, m') = A$ for all $m' \in M$.

In other words, Axiom P requires the message set M to contain at least one message for each action in the underlying G such that the action is exactly specified. To send such a message and then take another action violates the common understanding of the language, irrespective of the message sent by the other player. Such a message

could take the form "I will take action a irrespective of what message you send".[14] Likewise, Axiom N requires the message set to contain at least one message that does not specify what action in G the speaker will use, irrespective of what the other player says, for instance, "I promise nothing as to what action I will take, irrespective of what you say". Messages of the latter type will be called *null messages*.

Remark 1. *We obtain cheap talk as the special case when all messages are null messages ($\mu(m, m') = A$ for all $m, m' \in M$).*

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, and (iii) there is no other pure strategy that earns the same material payoff against σ and has a lower expected dishonesty cost. 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 $w(\sigma, \sigma) = 0$. We call such Nash equilibria *honesty equilibria*. By contrast, a symmetric Nash equilibrium (σ, σ) of $\tilde{\mathcal{G}}$ is a *dishonesty equilibrium* if $w(\sigma, \sigma) < 0$. The following example exhibits a dishonesty equilibrium.

Example 1 [Dishonesty equilibrium]**.** *Consider the game G defined by the payoff bimatrix in (1). Let $M = \{$"c", "d"$\}$, where "c" is honest iff c is played, $\mu($"c", $\cdot) \equiv \{c\}$, and "d" is honest iff d is played, $\mu($"d", $\cdot) \equiv \{d\}$.*[15] *Consider the pure strategy $s =$ ("d", f), where $f($"d", "d"$) = c$ and $f($"d", "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 material payoff. A deviation to "c" results in a material payoff loss, so (s, s) is also a Nash equilibrium in the lexicographic communication game $\tilde{\mathcal{G}}$, a dishonesty equilibrium.*

[14] Likewise, Rabin (1994), see section 5, 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 they want to suggest (op. cit. Definition 2).

[15] The message "c" could, for example, be "I will take action c" or "Let us play c".

Next, we consider the opposite possibility, discussed in Aumann (1990), namely that people may say "c" when they actually 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 any language in which (a) saying "c" is dishonest when followed by play of d, and (b) there is a message that is honest to send in conjunction with taking action d. More precisely, to send the message "c" with positive probability and then play the Nash equilibrium (d, d) of G in the preceding example is incompatible with Nash equilibrium in the lexicographic communication game.

Example 2 [Disequilibrium]**.** *Let $\tilde{\mathcal{G}}$ be as in the preceding example. Suppose that σ sends the message "c" with positive probability and that play of (σ, σ) results in the action pair (d, d) with probability one. Then each player incurs material payoff 7 and a positive expected dishonesty cost. A unilateral deviation to a pure strategy $s = (\text{"}d\text{"}, f)$, where $f(\text{"}d\text{"}, m) = d$ for all messages m, does not reduce the material payoff but reduces the dishonesty cost. Hence, (σ, σ) is not a Nash equilibrium of $\tilde{\mathcal{G}}$. By contrast, sending "d" with probability one and playing the action pair (d, d) is compatible with Nash equilibrium in $\tilde{\mathcal{G}}$.*

We now explore the implications of Axioms P and N. First, if the language contains a null message, then any symmetric Nash equilibrium of an underlying game 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 payoff dominated equilibrium (d, d) in the game in (1) is consistent with Nash equilibrium in $\tilde{\mathcal{G}}$. Denoting mixed strategies in G by $\rho \in \Delta(A)$, with $\rho(a)$ for the probability assigned to action $a \in A$, we have:

Lemma 2. *Let (ρ, ρ) be a Nash equilibrium of a symmetric two-player game G and suppose that $\tilde{\mathcal{G}}$ 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 $\rho(a)$.*

Second, if G is a coordination game with at least two actions, then every symmetric Nash equilibrium in the associated lexicographic communication game has a message that is not sent in equilibrium if axioms P and N are met. More precisely, we call a finite and symmetric $n \times n$-game G 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 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.

Lemma 3. *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.*

The following example shows that there are dishonest equilibria in some games even under the hypotheses of Lemma 3. It is as if two friends are joking with each other. They both say "let us meet at the bad restaurant" although they understand that the other actually plans to go to the good restaurant. A deviation from this joke would bewilder the other and induce him or her to indeed go to the bad restaurant.

Example 3 [Dishonesty despite Axiom N]**.** *Reconsider game G in (1) and let $M = \{$"c", "d", $n\}$, where "c" is honest if and only if c is played, "d" if and only if d is played, and n is a null message. Let $\tilde{\mathcal{G}}$ be the lexicographic communication game based on G, with message set M and the meaning correspondence described above; so $\tilde{\mathcal{G}}$ satisfies Axioms P and N. Consider the pure strategy $s = (\text{"}d\text{"}, f)$, where for all $m \in M$: $f(m, \text{"}d\text{"}) = c$ and $f(m, \text{"}c\text{"}) = f(m, n) = d$. In other words: say "d", and take action c if and only if you receive the message "d". Messages "c" and n 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.*

This example and Lemma 3 together show that, although a lexicographic preference for honesty does not rule out the possibility of lying equilibria, nevertheless it restricts the sets of messages sent in equilibrium. In particular, it rules out so-called babbling equilibria, that is, equilibria in cheap-talk games in which all messages are sent and "nobody listens" (actions are not conditioned on messages). This property is crucial for our main result.

The structure of the sets of Nash equilibria, in G and $\tilde{\mathcal{G}}$ respectively, are as follows. The cheap-talk game $\mathcal{G}$ is finite, so its Nash equilibria form a finite disjoint union of closed and connected semialgebraic sets, the Nash equilibrium components of $\mathcal{G}$. The same is true of the set Δ^{NE} of fixed points under β.[16] Likewise, the set $\tilde{\Delta}^{NE}$ can be defined in terms of finitely many real polynomial inequalities and so is a finite disjoint union of semialgebraic subsets of Δ^{NE}. It follows immediately from the definition of lexicographic Nash equilibria that each component (and hence its closure) of $\tilde{\Delta}^{NE}$ is contained in some component of Δ^{NE}. Components of $\tilde{\Delta}^{NE}$, unlike those of Δ^{NE}, need not be closed, due to the possibility of dishonesty equilibria. The next example illustrates this fact.

Example 4 [A non-closed component]**.** *Reconsider the strategy $s = (\text{"}d\text{"}, f)$ in Example 1, and let $s' = (\text{"}d\text{"}, f^c)$, where $f^c \equiv c$. Consider mixed strategies $\sigma_\lambda = \lambda s + (1 - \lambda) s'$, for $\lambda \in [0, 1]$. Note that $\sigma_\lambda \in \Delta^{NE}$ for all such λ. It is as if everybody says "d", plays c when hearing "d", and plays d with probability λ if someone would*

[16] This set is a projection of the intersection between the set of Nash equilibria in G and the diagonal of the space of mixed-strategy profiles. It is non-empty by Kakutani's Fixed-Point theorem applied to β, see Weibull (1995).

instead say "c". We also have $\sigma_\lambda \in \tilde{\Delta}^{NE}$ *for all* $\lambda > 0$*, but not for* $\lambda = 0$*. For positive* λ*, a deviation from message "d" to message "c" leads to a lower expected material payoff, since "c" is met by action d with positive probability. However, for* $\lambda = 0$ *a deviation to "c", followed by taking the action c, incurs no loss in expected material payoff (is not "punished") but raises the honesty payoff. Such a deviation is thus a lexicographically better reply to* σ_0*, and hence* $\sigma_0 \notin \tilde{\Delta}^{NE}$*. Thus the component of* $\tilde{\Delta}^{NE}$ *that contains the strategies* σ_λ*, for all* $\lambda > 0$*, is not closed.*[17]

4. 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; no small group of individuals in a large community can do better by together deviating to another strategy when the rest of the community plays the original strategy.[18] We here apply this concept to the *material* payoffs in lexicographic communication games (or equivalently, to the cheap-talk game associated with any given lexicographic communication game).[19] Formally,

Definition 1. *A mixed strategy* $\sigma \in \Delta(S)$ *is* ***neutrally stable*** *if* $\forall \tau \in \Delta(S)$*:*
(i) $v(\tau, \sigma) \leq v(\sigma, \sigma)$ *and*
(ii) $v(\tau, \sigma) = v(\sigma, \sigma) \;\Rightarrow\; v(\tau, \tau) \leq v(\sigma, \tau)$.

In other words, a neutrally stable strategy (an NSS) is a strategy σ that is a best reply to itself, in terms of material payoffs, and, in case of multiple best replies, fares at least as well against other best replies τ as these fare against themselves.[20] Let $\Delta^{NSS} \subset \Delta(S)$ denote the (closed but potentially empty) set of neutrally stable strategies. Clearly $\Delta^{NSS} \subset \Delta^{NE}$. We call a component X of $\tilde{\Delta}^{NE}$ *neutrally stable* if it is contained in Δ^{NSS}.

A closed set X of neutrally stable strategies is called evolutionarily stable (Thomas (1985)) if it contains all strategies τ that (a) are material best replies to some strategy

[17] In fact, honesty and closedness are strongly related properties: if axiom N and the game is not trivial components are closed if and only if they do not contain lying equilibria. The proof of this fact, that is not needed in the sequel is available upon request from the authors.

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

[19] Similar results are obtained if one includes honesty costs and applies evolutionary stability to the full lexiographic payoff structure.

[20] To see that this is equivalent with the above given verbal condition concerning post-mutation populations, note that, since v is linear in each of its two arguments, neutral stability is equivalent with requiring that for all τ: $v(\tau, (1-\varepsilon)\sigma + \varepsilon\tau) \leq v(\sigma, (1-\varepsilon)\sigma + \varepsilon\tau)$, for all $\varepsilon > 0$ small enough.

σ in X and (b) does materially just as well against themselves as strategy σ does against them:

Definition 2. *A non-empty and closed set $X \subset \Delta^{NSS}$ is **evolutionarily stable** if for all $\sigma \in X$ and $\tau \in \Delta(S)$:*

$$v(\tau,\sigma) = v(\sigma,\sigma) \wedge v(\tau,\tau) = v(\sigma,\tau) \quad \Rightarrow \quad \tau \in X. \tag{7}$$

Applied to a singleton set $\{\sigma\}$, this definition is identical with Maynard Smith's (1982) definition of an *evolutionarily stable strategy (ESS)* σ. Not surprisingly, evolutionarily stable sets thus have most of the properties of evolutionarily stable strategies. In particular, just as an ESS, viewed as a population state, is asymptotic stable in the replicator dynamic (Taylor and Jonker (1978)), an evolutionarily stable set is set-wise asymptotically stable in the same dynamic (Thomas (1985), Weibull (1995)). More precisely, if we view the probabilities assigned by a mixed strategy σ to pure strategies $s = (m, f)$ as population shares, in a population where individuals now and then are randomly pairwise matched to play the game $\tilde{\mathcal{G}}$, and if pure strategies that on average give higher material payoffs spread faster in the population than those that on average give lower material payoffs, then no small perturbation of a population state σ in, or near, an evolutionarily stable set X will lead the population state far away from X. Indeed, the population state will in the long run be arbitrarily close to, or in, the set X.[21] In this sense, (setwise and pointwise) evolutionary stability implies asymptotic stability in the replicator dynamic. It is also known that a neutrally stable strategy σ, again viewed as a population state, is weakly dynamically stable (or *Lyapunov stable*) in the replicator dynamic (Bomze and Weibull (1995)). It is easily verified that this also holds for any neutrally stable set.[22] Hence, no small perturbation of a neutrally stable population state, or a population state in or near a closed set of such states, will lead the population state far away.[23]

[21] A population state x, or, more generally, a compact set X of population states is *Lyapunov stable* if for every open set A containingat $x(0) \in B \Rightarrow x(t) \in A$ for all $t > 0$. In other words, starting in B, the population state will never leave A. A compact set X is *asymptotically stable* if it is Lyapunov stable and, moreover, there exists an open set B^* containing X such that $x(0) \in B^* \Rightarrow d(x(t), X) \to 0$. In other words, starting sufficiently near X, the population state will asymptotically approach X.

[22] To see this, let $X \subset \Delta^{NSS}$ be closed, and let $A \supset X$ be open. For each $x \in X$, let $x \in A_x$ for A_x open with $A_x \subset A$. There exists an open set B_x such that $x \in B_x$ and such that $x(0) \in B_x \Rightarrow x(t) \in A_x$ for all $t > 0$. The union $B = \cup_{x \in X} B_x$ is an open set that contains X, and $x(0) \in B \Rightarrow x(t) \in A_x \subset A$ for all $t > 0$.

[23] The closedness requirement is important, since each point in a non-closed set X can be Lyapunov stable and yet boundary points of X can be dynamically unstable. Binmore and Samuelson (1994), Binmore, Gale and Samuelson (1995), Weibull (1995) and Binmore and Samuelson (1999) analyze variants of entry-deterrence and ultimatum-bargaining games with precisely this property.

Let G be a finite and symmetric coordination game with a unique payoff-dominant Nash equilibrium (c, c), resulting in payoff α to each player.[24] For any real number β, call β a (material) *equilibrium outcome* in the lexicographic communication game $\tilde{\mathcal{G}}$ if $v(\sigma, \sigma) = \beta$ for some strategy σ in $\tilde{\Delta}^{NE}$, and say that a component X of $\tilde{\Delta}^{NE}$ *results* in payoff β if $v(\sigma, \sigma) = \beta$ for all strategies σ in X. Finally, call β an *evolutionarily stable outcome* if the set $X(\beta) = \{\sigma \in \Delta(S) : v(\sigma, \sigma) = \beta\}$ of strategies σ that earn material payoff β against themselves is evolutionarily stable.

We proceed to establish that if Axioms P and N are met, then an equilibrium component is neutrally stable if and only if it results in the payoff dominant outcome in the underlying game and, moreover, that this outcome is the unique evolutionarily stable equilibrium outcome. Formally:

Proposition 1. *Let G be a finite and symmetric coordination game with a unique payoff dominant Nash equilibrium with payoff α. Suppose that $\tilde{\mathcal{G}}$ is a lexicographic communication game, based on G, that satisfies Axioms P and N. A component X of $\tilde{\Delta}^{NE}$ is neutrally stable if and only if it results in material payoff α. Moreover, this outcome is the unique evolutionarily stable equilibrium outcome.*

While the proof given in the appendix is somewhat lengthy and technical, its intuition is simple. The most important claim is the instability of components X of $\tilde{\Delta}^{NE}$ that do not result in the maximal material payoff. Let X be such and suppose that $\sigma \in X$. By Lemma 3, there exists a message m that is not sent by σ. The population may drift in the component X towards strategies σ' that do not "punish" senders of m, and earn the same material payoff as σ (against σ and itself). This leaves the door open for mutants who use the message m as a "secret handshake" among themselves. They earn the same material payoff against σ' as the non-mutants do. However, by playing the action-pair (c, c) when meeting each other, they earn more in such encounters and thus also on average. The two parts of the argument, "drift to non-punishing strategies" and "secret handshake" are illustrated in the following two examples:

Example 5 [Drift]**.** *Let $\tilde{\mathcal{G}}$ be the lexicographic communication game in example 3. There is a Nash equilibrium in which both players say "hi" (send message n) to each other and then play the mixed Nash equilibrium in the underlying game G. More exactly, let $\sigma = \frac{7}{8}(n, f^*) + \frac{1}{8}(n, f^d)$, where $f^*(n, n) = c$, $f^*(n, \text{"c"}) = f^*(n, \text{"d"}) = d$ and $f^d(\cdot, \cdot) \equiv d$. It is easily verified that σ is a best reply to itself in $\tilde{\mathcal{G}}$, that is, $\sigma \in \tilde{\Delta}^{NE}$. Unilateral deviations to any other message are punished by play of d for sure, giving the deviator a material payoff of at most 7. However, also $\sigma' \in \tilde{\Delta}^{NE}$ for*

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

$\sigma' = \frac{7}{8}(n, f^c) + \frac{1}{8}(n, f^d)$, where $f^c(\cdot,\cdot) \equiv c$, a strategy that does not punish deviating messages. The two strategies σ and σ' belong to the same component of $\tilde{\Delta}^{NE}$, since $\sigma^t = (1-t)\sigma + t\sigma' \in \tilde{\Delta}^{NE}$ for all $0 \leq t \leq 1$. Hence, if strategies are subject to drift off their equilibrium paths, then drift may lead away from the punishing strategy σ to the "forgiving" strategy σ' in the same component of $\tilde{\Delta}^{NE}$.

Example 6 [Secret handshake]**.** *Let again $\tilde{\mathcal{G}}$ be the lexicographic communication game in example 3 and now let $\sigma \in \tilde{\Delta}^{NE}$ be such that $v(\sigma,\sigma) = 7$, that is, σ plays d against itself. By Lemma 3, there exists a message m that is not sent by σ. Let τ send m, play c when receiving m, otherwise d. Then $v(\tau,\sigma) \geq 7 = v(\sigma,\sigma)$. Moreover, $v(\tau,\tau) = 9 > v(\sigma,\tau)$, where the last inequality holds since σ has to play d against τ; otherwise there would exist a profitable unilateral deviation against σ in $\mathcal{G}$. This proves that σ is not neutrally stable, $\sigma \notin \Delta^{NSS}$. A small group of mutants playing τ would do better than σ in the post-entry population. Consequently, the component of $\tilde{\Delta}^{NE}$ to which σ belongs is not even weakly evolutionarily stable.*

5. Sender-receiver games

We here briefly discuss how our approach can be extended to games with one-sided communication—so-called sender-receiver games—and what results the approach yields. Intuitively, one would expect one-sided communication to be beneficial for the sender, who arguably can lead play towards any preferred Nash equilibrium in the underlying game G. This intuition turns out to be roughly, though not entirely, right.

Let G be a finite, not necessarily symmetric, two-player game with player roles S and R and with action set A for player S and action set B for player R. Let the payoffs to the pure-strategy pair $(a,b) \in A \times B$ be $\pi_S(a,b)$ and $\pi_R(a,b)$. Define a cheap-talk sender-receiver game $\mathcal{H}$, based on G, as follows. Before G is played, S sends a message m from a finite set M. Player R receives this message and thereafter both players simultaneously take their actions, $a \in A$ and $b \in B$, respectively, in game G. Hence, in $\mathcal{H}$ a pure strategy for the sender is a pair $(m,a) \in M \times A$ and, for the receiver, a function $g : M \to B$ that maps received messages to own actions. Play of such a pure-strategy pair in $\mathcal{H}$ results in actions a and $b = g(m) \in B$ in G. The payoffs in $\mathcal{H}$ are the resulting payoffs in the underlying game G, to be called the material payoffs.

We introduce meaning of messages in a similar way as in games with two-sided communication, and thereby obtain a game $\tilde{\mathcal{H}}$ with one-sided communication. More specifically, for each message $m \in M$, let $\mu(m)$ be a subset of the sender's action set A. The elements of $\mu(m)$ are those actions that message m means that the sender intends to play. This defines the meaning correspondence $\mu : M \rightrightarrows A$ in $\tilde{\mathcal{H}}$. Let $h : M \times A \to \mathbb{R}_+$ be an function such that $h(m,a) = 0$ if and only if $a \in \mu(m)$; this

is the sender's honesty cost function. Define the sender's lexicographic preferences in $\tilde{\mathcal{H}}$ along the same lines as in games with two-sided communication, and let the receiver's preferences in $\tilde{\mathcal{H}}$ simply be defined by this player's material payoffs. This defines $\tilde{\mathcal{H}}$ as a lexicographic sender-receiver game. Now define $\bar{\mathcal{H}}$ as the symmetric lexicographic communication game that is obtained from $\tilde{\mathcal{H}}$ by adding a first random draw by "nature", whereby one of the players in $\bar{\mathcal{H}}$ becomes the sender and the other the receiver in $\tilde{\mathcal{H}}$, with equal probability for both draws. Both players in the symmetric game $\bar{\mathcal{H}}$ thus have a lexicographic preference for honesty, a preference that matters only if the player happens to be drawn to be the sender.

Consider lexicographic communication games $\bar{\mathcal{H}}$ satisfying axioms P and N.[25] For each mixed-strategy profile (σ, τ), denote by $v_S(\sigma, \tau)$ and $v_R(\sigma, \tau)$ the *conditionally expected* material payoffs to the player who plays σ, conditional upon the event that this player is drawn to be the sender and receiver, respectively. Thus

$$\bar{v}(\sigma, \tau) = \frac{1}{2} v_S(\sigma, \tau) + \frac{1}{2} v_R(\sigma, \tau) \tag{8}$$

is the expected material payoff to strategy σ in $\bar{\mathcal{H}}$ when played against τ. Let $\bar{\beta}$ be the best-reply correspondence in $\bar{\mathcal{H}}$ and let $\bar{\Delta}^{NE}$ be its set of fixed points. Again, this set consists of finitely many connected components.

We establish the claimed results for a class of games G that contain those considered in Proposition 1 — symmetric coordination games with a unique payoff dominant Nash equilibrium. Generalizing the notation in the preceding section somewhat, let α_S denote the maximal payoff in G to player role S,

$$\alpha_S = \max_{(a,b) \in A \times B} \pi_S(a, b). \tag{9}$$

In other words, whenever the player in the sender role obtains this payoff, he gets "his way" in G. For any real number β, call β an *equilibrium sender-outcome* if $v_S(\sigma, \sigma) = \beta$ for some strategy σ in $\bar{\Delta}^{NE}$, and say that a component X of $\bar{\Delta}^{NE}$ *results* in sender-payoff β if $v_S(\sigma, \sigma) = \beta$ for all strategies σ in X. Finally, call β an *evolutionarily stable sender-outcome* if the set $X_S(\beta) = \{\sigma \in \Delta(S) : v_S(\sigma, \sigma) = \beta\}$ is evolutionarily stable.

Consider games G in which the maximum payoff α_S to player role S is achieved in only *one* action pair, (a^*, b^*), and, moreover, this action pair is a *strict* Nash equilibrium of G. We call such games *strict*. In such a game, a message suggesting play of (a^*, b^*) is *self-committing* in the sense that a sender believing that the receiver believes the message has an incentive to carry out her action, a^*. However, such a

[25] In this context, axiom P requires that there for each $a \in A$ exists at least one $m \in M$ such that $\mu(m) = \{a\}$, where μ is the meaning correspondence in the associated game $\tilde{\mathcal{H}}$. Likewise, axiom N requires that there exists at least one message $n \in M$ such that $\mu(n) = A$.

message need not be self-signaling, where *self-signaling* means that the sender prefers the receiver to believe the message *only* if she plans to carry out her action. Aumann's example shows that strictness does not imply this property. Clearly all symmetric coordination games with a unique payoff-dominant Nash equilibrium are strict in this sense, as are all games of the battle-of-sexes and hawk-dove varieties (while prisoner's dilemma games are not). One would guess that, when the underlying game is strict, it would be evolutionarily advantageous to "declare to be tough" and then play a^* when in the sender role, and to be accommodating and play b^* in the receiver role. The following proposition formalizes this intuition.[26]

Proposition 2. *Let G be strict and let $\bar{\mathcal{H}}$ be a symmetric lexicographic communication game, based on G, that satisfies axioms P and N. A component X of $\bar{\Delta}^{NE}$ is neutrally stable if and only if it results in the maximal sender-payoff α_S. Moreover, this is the unique evolutionarily stable equilibrium sender-outcome.*

The intuition for the proof, put in the appendix, is as follows. First, in a Nash equilibrium there is always an unused message, as in the case of two-sided communication. Second, the population may drift, within the component in question, towards a strategy that does not "punish" senders of the unused message. Third, if such a "forgiving" strategy does not induce the maximal sender-payoff, this leaves the door open for mutants who "get their way" as senders and accommodate optimally as receivers. On average, such mutants earn a higher material payoff than the rest of the population.

Unlike in the case of two-sided communication, the unique evolutionary stable outcome need not be *ex-ante* Pareto efficient. To see this, let G be the following skewed battle-of-the-sexes game:

$$\begin{array}{cc} c & d \\ c \;\; 3,1 & 0,0 \\ d \;\; 0,0 & 2,6 \end{array} \tag{10}$$

This game has two strict Nash equilibria, one better for the row player, the other better for the column player, both Pareto efficient in the game G, but the latter giving the highest average payoff, 4. The game G is clearly strict, with $\alpha_S = 3$. Hence, in a lexicographic and symmetrized communication game $\bar{\mathcal{H}}$ satisfying axioms P and N, the unique evolutionarily stable set results in play of the strict equilibrium (c, c) preferred by the player in the sender role, although the associated expected material payoff in $\bar{\mathcal{H}}$ is only 2, while always sending a null message and taking action d is

[26] The conclusions are valid under less stringent, but more involved assumptions on G than strictness.

another Nash equilibrium of $\bar{\mathcal{H}}$ and results in the higher expected material payoff 4. Why cannot a few mutants playing (d, d) with each other invade a population playing (c, c)? The point is that, even if such mutants would appear and start sending an unused message (in order to recognize each other), and even if the rest of the population would not punish senders of this message, the mutant in the sender role has no way to know whether the receiver is a mutant or not, where the latter is much more likely. So a mutant sender essentially has to presume that the receiver is a non-mutant and will therefore have no way to be "nicer" to a receiving mutant. By contrast, under two-sided communication, two mutants, both sending an unused message, can recognize each other and thereby coordinate their actions on a better equilibrium in the underlying game.[27]

6. Related work

Intuition and experiments suggest that pairs of individuals are usually able to achieve efficiency in coordination games when they are allowed to communicate before play. Moreover, the ability to coordinate seems to be greater the closer to a natural language the experimental communication protocol is, see Valley *et al.* (2002) and Charness and Dufwenberg (2006). We here model a shared language, along with the cultural conventions in its use, by way of a meaning correspondence. In this we differ from the cheap-talk literature and from other models of pre-play communication. This section briefly comments on some of the most closely related work, in chronological order.

Farrell (1988,1993) analyzes costless pre-play communication when messages have a pre-existing meaning. Unlike here, players have no preference for honesty *per se*. Instead, Farrell imposes a credibility condition, roughly requiring the listener to believe the speaker unless the speaker could have a "strategic reason" to mislead the listener. Credibility is a property of a message (and may depend on the game in question), while we model honesty as a property of a triplet—a message-pair and an action—and assume that players have a deontological preference for this property.

Myerson (1989) develops a formal credibility criterion for one-sided communication games, assuming that messages have a pre-existing meaning. Applied to games of complete information, Myerson's criterion essentially requires that if the sender promises to take a certain action and recommends the others to take some actions, the so defined action-profile should constitute a Nash equilibrium of the underlying game. Players do not have deontological preferences for honesty or the truth.

As mentioned in the introduction, Robson (1990) pioneered the idea of using unsent messages as "secret handshakes" among mutants (see also Wärneryd (1991)). Using a similar argument, Sobel (1993) establishes a form of dynamic evolutionary

[27] The line of reasoning in this section applies also to cheap-talk games, at the expense of introducing one more round of evolutionary drift in order to create an unused message. In that setting, our result essentially replicates Proposition 4 in Kim and Sobel (1995).

stability of efficient outcomes in coordination games preceded by two-sided cheap talk. He defines a population dynamic with a finite population for each player role in a two-player game. Pairs of individuals, one from each player population, are randomly matched to play the game and all individuals play pure strategies. Sobel assumes that there are more messages than individuals in each player population, so there always exists at least one unsent message. Evolutionary drift may lead to a population state in which the unsent message in question is not "punished". If, in a coordination game, the average population payoff is not maximal, this opens the door for mutants to destabilize the population state by way of sending the unsent message and playing the "good" Nash equilibrium among themselves.[28]

Sobel's (1993) model is further developed in Kim and Sobel (1995) and they also consider sender-receiver games. As pointed out earlier, our result for this case confirms theirs, the only difference being that our approach requires one round less of evolutionary drift. This suggests the possibility that a honesty costs might induce faster convergence to the equilibrium preferred by the sender.

Rabin (1994) analyzes two-sided pre-play communication in symmetric two-player games. He considers costless communication in a language with pre-existing meaning, and players make repeated simultaneous proposals before playing the underlying game. If all players propose the same equilibrium in a given pre-play communication round, then this is taken to be an agreement to play that equilibrium. Our approaches differ, since in Rabin's model players do not have honesty preferences and in our model two identical messages are not taken to necessarily constitute an agreement.

Blume (1998) studies a stochastic population dynamic for pre-play communication games in which some messages have *a priori* meaning. Namely, 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 when playing according to that equilibrium, while the sender's payoff is unaffected. By contrast, we assume that it is the sender who may incur a lexicographic payoff loss, while the receiver's payoff does not depend directly on the message received.

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 communication stage, that is, of not sending a message and not knowing if the other player has sent a message. By contrast, while our null axiom permits senders to avoid "commit-

[28] As showed by Schlag (1993, 1994), Wärneryd (1998) and Banerjee and Weibull (1993, 2000), this argument does not apply if individuals are allowed to play mixed strategies and the game in question is played by individuals drawn from one and the same population. In particular, there exists an evolutionarily stable outcome in 2×2 coordination games that sends all messages and results in suboptimal payoffs.

ting" to any particular action, receivers in our model cannot commit to not observe the other's message. Hurkens and Schlag show that in their setting the unique evolutionarily stable set in $n \times n$-coordination games is characterized by play of the payoff dominant equilibrium.

Kartik, Ottaviani and Squintani (2007) develop a sender-receiver model of the Crawford and Sobel (1982) variety and use this to analyze strategic misrepresentation of private information. The sender knows the true state and incurs a disutility from misrepresenting his private information. The message space is identical with the state space and this is an unbounded interval. The sender's utility function is decreasing in his message's deviation from the truth. Receivers have a small probability of being credulous. See Chen (2004), Kartik (2005) and Chen, Kartik and Sobel (2007) for more research on this topic.[29]

Lo (2007) develops an alternative model of language, meaning and games. She focuses on sender-receiver games and formalizes meaning by way of restricting the receiver's reactions to messages. Her solution concept is iterated elimination of weakly dominated strategies. In battle-of-the-sexes games, the sender obtains her preferred outcome, as in our model (Section 5), while all outcomes are possible in Aumann's example (1), in sharp contrast with our model.

7. Concluding comments

An interesting feature of evolutionary stability in pre-play communication games is its logical independence of ordinality in the underlying game, where by ordinality we mean invariance of the solution under transformations that leave the best-reply correspondence unchanged. For example, while the best-reply correspondence of the game in (1) is identical with that of

$$\begin{array}{cc|cc} & & c & d \\ c & & 1,1 & 0,0 \\ d & & 0,0 & 7,7 \end{array} \tag{11}$$

the unique evolutionarily stable outcome in a lexicographic pre-play communication game, as modelled above, is play of (c,c) when based on (1) but (d,d) when based on (11).[30] A more profound question, falling outside the scope of this study, is whether indeed ordinality should be viewed as a general desideratum for solution concepts in games.

[29] For other analyses of deceit and lying, see Sobel (1985), Benabou and Laroque (1992), Farrell and Gibbons (1989), Conlisk (2001), Crawford (2003) and Miettinen (2006).

[30] 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 in the game to which they are applied, here the cheap-talk game $\mathcal{G}$.

We plan to extend our analysis in different directions. We plan to apply this approach to infinitely repeated games. Fudenberg and Maskin (1991) showed that evolutionary stability and noise together have strong efficiency implications in infinitely repeated prisoners' dilemmas. If, instead of noise, the players communicate with each other between rounds, will this destabilize inefficient outcomes? We also intend to study the evolution of meaning and honesty of language in populations using cheap talk.

8. Appendix

This section contains mathematical proofs of results not proved in the main text.

8.1. Lemma 3. 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 n be such a message. Since n 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 $f(m, m') \in \mu(m, m')$ 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 $\mu(\text{"}c\text{"}, \cdot) \equiv \{c\}$ and $\mu(\text{"}d\text{"}, \cdot) \equiv \{d\}$. Since, by hypothesis, every message is sent in σ, the message pair ("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, $\sigma \notin \tilde{\beta}(\sigma)$.

8.2. Proposition 1. First, we prove that a component X of $\tilde{\Delta}^{NE}$ is not neutrally stable if it contains a strategy σ with $v(\sigma, \sigma) < \alpha$.

Since $\sigma \in \tilde{\Delta}^{NE}$ and axioms N and P hold, Lemma 3 implies that there exists at least one message that is not sent in σ. Let $n \in M$ be such. Choose a pure strategy $\tilde{s} \in supp(\sigma)$ such that $v(\sigma, \tilde{s}) < \alpha$ and let $\tilde{s} = (\tilde{m}, \tilde{f})$. For each pure strategy $s = (m, f) \in supp(\sigma)$, let s^1 be the associated modified pure strategy (m, f^1), where $f^1(m, m') = f(m, m')$ for all $m' \neq n$ and $f^1(m, n) = f(m, \tilde{m})$. Note that s^1 reacts to receiving n just as s reacts to $\tilde{m}$, while otherwise they coincide. If $\sigma = \sum \lambda_i s_i$ for pure strategies s_i and probability weights $\lambda_i > 0$ summing to 1, let $\sigma^1 = \sum \lambda_i s_i^1$. In other words, σ^1 is the same convex combination of the pure strategies s_i^1 as σ is with respect to the s_i. For all $0 \leq t \leq 1$, define $\sigma^t = (1-t)\sigma + t\sigma^1$. Note that σ and σ^t send the same messages and react in the same way to the messages they send: they differ only in their reaction to the (unsent) message n. This implies that $v(\sigma, \sigma) = v(\sigma^t, \sigma^t)$ and $w(\sigma, \sigma) = w(\sigma^t, \sigma^t)$ for all $0 \leq t \leq 1$.

$\sigma^t \in \tilde{\Delta}^{NE}$ for all $t < 1$.

Proof: Suppose first that $\tau \in \Delta(S)$ does not use n. Then $v(\tau, \sigma^t) = v(\tau, \sigma) \leq v(\sigma, \sigma) = v(\sigma^t, \sigma^t)$ since $\sigma \in \tilde{\Delta}^{NE}$. Under equality, $w(\tau, \sigma^t) = w(\tau, \sigma) \leq w(\sigma, \sigma) = w(\sigma^t, \sigma^t)$ for the same reason. Suppose now that $\tau \in \Delta(S)$ uses n**.** Consider thus a

pure strategy sending n, say $s^0 = (n, h)$. Clearly $v(s^0, \sigma) \leq v(\sigma, \sigma)$, because $\sigma \in \tilde{\Delta}^{NE}$. Moreover, by definition of σ^1, we have $v(s^0, \sigma^1) = v((\tilde{m}, h), \sigma)$. Now, the fact that $\sigma \in \tilde{\Delta}^{NE}$ implies that

$$v((\tilde{m}, h), \sigma) \leq v(\sigma, \sigma) = v(\sigma^1, \sigma^1).$$

If $v(s^0, \sigma) < v(\sigma, \sigma)$, linearity in t implies that $v(s^0, \sigma^t) < v(\sigma^t, \sigma^t)$ for all $t < 1$. If $v(s^0, \sigma) = v(\sigma, \sigma)$, then we must have $w(s^0, \sigma) \leq w(\sigma, \sigma)$ because σ is a lexicographic NE. However, we also have $w(s^0, \sigma^1) = w(s^0, \sigma)$ because, again, σ and σ^1 send the same messages, so

$$w(s^0, \sigma^1) = w(s^0, \sigma) \leq w(\sigma, \sigma) = w(\sigma^1, \sigma^1),$$

Again, by linearity in t, we are done. This proves the claim.[31]

Consider now the pure "secret handshake" strategy $\hat{s} = (n, g)$, where $g(n, m') = \tilde{f}(\tilde{m}, m')$ if $m' \neq n$ and $g(n, n) = c$. We then have $v(\hat{s}, \sigma^1) = v(\tilde{s}, \sigma) = v(\sigma, \sigma) = v(\sigma^1, \sigma^1)$ and $v(\hat{s}, \hat{s}) = \alpha$. However, $v(\sigma^1, \hat{s}) = v(\sigma, \tilde{s}) < \alpha$. Thus $\sigma^1 \notin \Delta^{NSS}$. But $\sigma^1 \in \bar{X}$ and $\bar{X} \subset \Delta^{NSS}$ since Δ^{NSS} is closed, establishing that X is not neutrally stable.

Secondly, we prove that α is the unique evolutionarily stable equilibrium outcome.

For any $\sigma \in X(\alpha)$ and $\tau \in \Delta(S)$, $v(\tau, \sigma) \leq \alpha = v(\sigma, \sigma)$, so condition (i) in the definition of neutral stability holds. If $v(\tau, \sigma) = v(\sigma, \sigma)$, then τ must always play c against σ, so $v(\sigma, \tau) = \alpha \geq v(\tau, \tau)$ and thus also condition (ii) in the definition of neutral stability holds. In sum: $X(\alpha) \subset \Delta^{NSS}$. We note that $X(\alpha)$ is non-empty (send any null message and react to all messages by taking action c) and that it is closed by continuity of v. It thus only remains to verify that if $v(\sigma, \sigma) = \alpha$, $v(\tau, \sigma) = v(\sigma, \sigma)$ and $v(\tau, \tau) = v(\sigma, \tau)$, then also $v(\tau, \tau) = \alpha$. But this follows from the above observation that $v(\sigma, \tau) = \alpha$.

Thirdly, it remains to prove that a component X of $\tilde{\Delta}^{NE}$ is neutrally stable if it results in material payoff α. However, this follows directly from the just proved fact that α is an evolutionarily stable equilibrium outcome, which implies that any subset of $X(\alpha)$ consists of neutrally stable strategies.

8.3. Proposition 2. We first prove that a component X of $\bar{\Delta}^{NE}$ is not neutrally stable if it contains some strategy σ such that $v_S(\sigma, \sigma) < \alpha_S$.

Let $\sigma \in X$. Using the same argument as in the proof of Lemma 3, it is not difficult to show that there exists a message $m \in M$ such that σ assigns zero probability to all pure strategies using m. We define σ^1 in a similar way as in the proof of Proposition

[31] Note that, if σ in a dishonest component, σ^1 does not necessarily belong to $\tilde{\Delta}^{NE}$ because we could have $v((n, h), \sigma) < v(\sigma, \sigma)$, $v((n, h), \sigma^1) = v(\sigma^1, \sigma^1)$ and $w((n, h), \sigma^1) > w(\sigma^1, \sigma^1)$, see Example 4.

1: in the sender role, σ^1 sends the same messages and takes the same actions as σ. In the receiver role σ^1 reacts to all messages but m in the same way as σ does. When m is sent, σ^1 reacts by taking action b^*. Let τ be the strategy that, in the sender role, sends message m and takes action a^*, and, in the receiver role behaves like σ^1. For $t \in [0,1]$, let $\sigma^t = (1-t)\sigma + t\sigma^1$. Suppose that $v_S(\sigma,\sigma) = \beta < \alpha_S$. It is easy to see that there is a $\bar{t} < 1$ such that, for all $t \in [0,\bar{t}]$, $\sigma^t \in X$. Let $\sigma' = (1-\bar{t})\sigma + \bar{t}\sigma^1$. Then $v_S(\tau,\sigma') = v_S(\sigma',\sigma') = \beta$. We claim that σ' is not neutrally stable, because τ is a successful mutant against it. First, τ is a material best reply to σ' in $\bar{\mathcal{H}}$. Because $v_R(\tau,\sigma') = v_R(\sigma',\sigma')$ since τ, in the receiver role, reacts to σ' exactly as σ' does, and τ does just as well as σ' in the receiver role. Hence, $\bar{v}(\tau,\sigma') = \bar{v}(\sigma',\sigma')$. Second, τ earns more material payoff against itself than σ' earns against it. To see this, first note that $v_S(\tau,\tau) = \alpha_S$ while $v_S(\sigma',\tau) = v_S(\sigma',\sigma') = \beta$, so $v_S(\tau,\tau) > v_S(\sigma',\tau)$. Secondly, $v_R(\tau,\tau) = \pi(b^*,a^*)$ while $v_R(\sigma',\tau) \leq \pi(b^*,a^*)$ because (a^*,b^*) is a Nash equilibrium. Hence, $\bar{v}(\tau,\tau) > \bar{v}(\sigma',\tau)$. Consequently, σ' is not neutrally stable and therefore X is not a neutrally stable component either.

Secondly, we prove that α_S is the unique evolutionarily stable equilibrium sender-outcome.

Let $\sigma \in X_S(\alpha_S)$. Since the maximum of v_S is attained by play of the unique action pair (a^*,b^*) in G, (σ,σ) induces play of (a^*,b^*) with probability one. In particular, σ always plays a^* in the sender role. Let τ be any mutant. By definition of $X_S(\alpha_S)$, $v_S(\tau,\sigma) \leq v_S(\sigma,\sigma)$. When τ is a receiver and σ the sender, σ takes action a^* and τ thus gets at most, $\pi(b^*,a^*)$, the payoff to the best reply to a^*. So $v_R(\tau,\sigma) \leq v_R(\sigma,\sigma)$. The same argument, and the strictness of the Nash equilibrium (a^*,b^*) in G, implies that if $\bar{v}(\tau,\sigma) = \bar{v}(\sigma,\sigma)$ then $v_S(\tau,\sigma) = v_S(\sigma,\sigma)$, $v_R(\tau,\sigma) = v_R(\sigma,\sigma)$ and the action pair (a^*,b^*) is played with probability one when τ meets σ. The latter implies that $\bar{v}(\sigma,\tau) = \bar{v}(\sigma,\sigma)$ and that τ always takes action a^* in the sender role. But then the best τ can do upon meeting itself is to play b^* in the receiver role, so $v(\tau,\tau) \leq v(\sigma,\tau)$, with equality implying that $\tau \in X_S(\alpha_S)$.

Thirdly, it remains to prove that a component X of $\bar{\Delta}^{NE}$ is neutrally stable if it results in the maximal sender-payoff α_S. However, this follows directly from the just proved fact that α_S is an evolutionarily stable equilibrium sender-outcome, which implies that any subset of $X_S(\alpha_S)$ consists of neutrally stable strategies.

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 (2006): "Screening ethics when honest agents care about fairness", *International Economic Review* 47, 59-85.

[4] Alger I. and R. Renault (2007): "Screening ethics when honest agents keep their word", *Economic Theory* 30, 291-311.

[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 (1993): "Evolutionary selection with discriminating players", Harvard University WP 1616.

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

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

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

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

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

[12] Binmore K., D. Gale and L. Samuelson (1995): "Learning to be imperfect: the ultimatum bargaining game", *Games and Economic Behavior* 8, 156-190.

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

[14] Binmore K. and L. Samuelson (199): "Evolutionary drift and equilibrium selection", *Review of Economic Studies* 66, 363-393.

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

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

[17] 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*.

[18] Bomze I. and J. Weibull (1995): "Does neutral stability imply Lyapunov stability?", *Games and Economic Behavior* 11, 173-192.

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

[20] Charness G. and M. Dufwenberg (2006): "Promises and partnership", *Econometrica* 74, 1579 - 1601.

[21] Chen Y. (2004): "Perturbed communication games with honest senders and naïve receivers", mimeo., Yale University.

[22] Chen Y., N. Kartik and J. Sobel (2007): "On the robustness of informative cheap talk", *Econometrica*, forthcoming.

[23] 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.

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

[25] 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.

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

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

[28] Crawford V. and J. Sobel (1982): "Strategic information transmission", *Econometrica* 50, 1982, 1431-1452.

[29] Davidson R. and K. Hugdahl (1995): *Brain Asymmetry*. Cambridge, MA: MIT Press.

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

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

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

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

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

[35] Gilboa I., Matsui A. (1991): "Social stability and equilibrium", *Econometrica* 59, 859-867.

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

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

[38] 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.

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

[40] Kartik N. (2005): "Information transmission with almost-cheap talk", mimeo., University of California at San Diego.

[41] Kartik N., M. Ottaviani and F. Squintani (2007): "Credulity, lies, and costly talk", *Journal of Economic Theory* 134,93-116.

[42] Kim Y.-G. and J. Sobel (1995): "An evolutionary approach to pre-play communication", *Econometrica* 63, 1185-1193.

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

[44] Kozel F., T. Padgett and M. George (2004): "A replication study of the neural correlates of deception", *Behavioral Neuroscience* 118, 852-856.

[45] Lundquist T., T. Ellingsen, E. Gribbe and M. Johannesson (2007): "The cost of lying", SSE/EFI Working Papers in Economics and Finance, No. 666.

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

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

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

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

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

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

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

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

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

[55] Schlag K. (1993): "Cheap talk and evolutionary dynamics", Bonn Department Discussion Paper B-242.

[56] Schlag K. (1994): "When does evolution lead to efficiency in communication games?", Bonn University Discussion Paper B-299.

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

[58] Sobel J. (1993): "Evolutionary stability and efficiency", *Economics Letters* 42, 301-312.

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

[60] Valley K., L. Thompson, R. Gibbons and M. Bazerman (2002): "How communication improves efficiency in bargaining games", *Games and Economic Behavior* 38, 127-155.

[61] Wärneryd K. (1991): "Evolutionary stability in unanimity games with cheap talk", *Economics Letters* 36, 375-378.

[62] Wärneryd K. (1998): "Communication, complexity, and evolutionary stability", *International Journal of Game Theory* 27, 599-609.

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