Commitments, Intentions, Truth and Nash Equilibria[*]

Karl H. Schlag[†] and Péter Vida[‡]

February 13, 2013

Abstract

Games with multiple Nash equilibria are believed to be easier to play if players can communicate. We present a simple model of communication in games and investigate the importance of when communication takes place. Sending a message before play captures talk about intentions, after play captures talk about past commitments. We focus on equilibria where messages are believed whenever possible. Applying our results to Aumann's Stag Hunt game we find that communication is useless if talk is about commitments, while the efficient outcome is selected if talk is about intentions. This confirms intuition and empirical findings in the literature.

Keywords: Pre-play communication, cheap talk, coordination.

JEL Classification Numbers: C72, D83.

[*] We would thank Olivier Gossner, Michael Greinecker, Toomas Hinnosaar, Melody Lo and Joel Sobel.

[†] University of Vienna, Department of Economics. E-mail: karl.schlag@univie.ac.at

[‡] University of Mannheim, Department of Economics. E-mail: vidapet@gmail.com

1 Introduction

Game theory is agnostic about how to play in games that have multiple Nash equilibria. Beliefs can be mutually self-confirming when all believe that others focus on an inefficient equilibrium even if there are alternative Nash equilibria where all are strictly better off. Yet, it is commonly believed that inefficient equilibria will not be played when players are allowed to communicate before they play the game. The reasoning is that it suffices that one player proposes an equilibrium outcome in which all players are better off to upset beliefs associated to inefficient play.

At the same time Aumann (1990) claims that communication can be useless even in the simplest games, and illustrates this informally in a version of the Stag Hunt game. Farrell (1988)[1] objects and argues for this game that it depends on when communication takes place. If communication occurs after the person communicating has made a choice then he agrees. However, if communication occurs before making a choice then he argues that communication will lead all players to hunt the stag. Charness (2000) runs experiments for this game that reinforce the intuition of Farrell.

We present a simple formal framework to examine credible communication, where players are believed whenever possible, in two person normal form games. We believe this to be a first step towards a general framework for analyzing credible communication with many players and incomplete information.

Simply adding cheap talk will not reduce the set of equilibrium outcomes. A necessary condition for upsetting beliefs supporting an inefficient equilibrium is that alternative proposals can be made. These would be initiated by sending unanticipated messages, naturally accompanied by an explanation of the circumstances surrounding the new proposal. One would also explain

[1] This is based on earlier personal communication on this matter, see Farell (1988).

which messages one would have sent if one had other intentions or the circumstances would be different. For communication to then be successful the parties involved, both those that talk and those that listen, must be able and willing to rethink their intentions.

We embed these ingredients into a standard game theory analysis, by setting up the rules of communication and adding features to the strategic interaction that capture what happens when one explains behavior under alternative circumstances. In our analysis we then only consider those equilibria where messages are believed whenever possible.

We model communication about play in a two player normal form game by letting player one (the sender) send a message to player two (the receiver). A message is embedded in a language that specifies the different possible statements. We enrich the game by allowing one of the two players to chose the language. Thereby we can investigate the importance of choosing the rules that govern communication. Whether or not players are able or willing to change their behavior based on what has been communicated plays an important role when communicating. This is captured in two extreme scenarios. In the scenario we call "first talk then play" (TP), communication occurs before either player has chosen an action. Player one as the sender is given the possibility to talk about which action she intends to choose later in the underlying normal form game. To model player one also talking about why she did not choose one of other messages, we add for each message a small probability that player one is forced to send this message. Player two as the receiver does not know whether or not the message has been sent freely. In the other scenario called "first play then talk" (PT), communication only takes place after (both know that) the player one has already chosen an action. Here player one as sender is given the possibility to talk about which action she has chosen. In order to model talk about what she would say if she had chosen a different action, we add for each action a small probability that player one is forced to choose this action. Player two as receiver does not know whether or not player one was free to choose the action. This second scenario can also be interpreted as common knowledge that the sender is unwilling to change her intentions.

Notice that our modeling of communication in games follows the two alternative settings mentioned by Farrell (1988). In “first play then talk” (PT) the person communicating has already chosen an action, while in “first talk then play” (TP) communication occurs before any actions have been chosen.

A key innovation is the use of perturbations to incorporate in equilibrium the consequences of unintended choices. Player one has to face the consequences of having sent an unintended message in TP and of having chosen an unintended action in PT. These perturbations are not mistakes but a way to incorporate alternative scenarios and counterfactuals in our model, features that play an important role in communication. After all, messages only obtain meaning by the context where they are used and when they are not used. Yet we did not want to blur the analysis by some sophisticated refinement or complicate the game by modelling how player one can justify her behavior.

Communication allows the sender to inform the receiver about intentions in TP and about past choices (or unrevocable commitments) in PT. In TP the message of player one could be an indication of how she intends to play in the game. E.g., “I will hunt the stag”. In PT player one could be providing information to player two about which action she has chosen. E.g., “I have already committed to hunt the stag”. The natural way of passing on information through communication is to tell the truth. Whether or not this is possible depends both on the beliefs of the receiver, whether she believes that player one is truthful, and on the incentives of the sender, whether it is best to tell the truth. It also depends on the language. For instance, if player one says "I will hunt either the stag or a rabbit" then player one can be truthful and yet no information will be transmitted if player one always makes this statement.

The only role of messages is to allow player two, given her beliefs, to differentiate between the different possible intentions in TP, and previous choices in PT, of player one. Hence, we restrict attention to messages that belong to a partition of the action space of player one. We call such a partition a language. In Aumann's Stag Hunt game there are two possible languages. {{Stag, Rabbit}} is the degenerate language that contains a single message. With this language information cannot be transmitted, it is as if players do not communicate. {{Stag}, {Rabbit}} is the language where player one can reveal her true intentions in TP and her true choice in PT. The language is chosen right before player one sends a message. In particular this means that the language is chosen in PT after player one has made a choice. Typically we imagine that player one chooses the language. "Look, I am telling you the truth about everything I could tell you about." implies that player one is choosing {{Stag}, {Rabbit}} as language. However, we also consider the situation where player two chooses the language, to separate incentives to tell the truth from the incentives to manipulate the context. The player who is assigned to choose the language is called the *interpreter*.

In the story behind the folklore that communication leads to efficiency, the sender needs to be able to convince the receiver that she wants to choose something different. We postulate that the receiver will believe the sender whenever all messages in the chosen language can be believed. A language in which all messages can be believed is called *credible*. Trivially, the degenerate language with a single message is credible. If the language is not credible then we postulate that the receiver ignores the message and acts as she does under the degenerate language. Thus, the sender can convince the receiver if she chooses a message from a credible language. Let us illustrate how this influences play in Aumann's Stag Hunt game. Consider TP with player one as interpreter. Then it cannot happen that no information is transmitted and both believe the other will hunt the rabbit. Namely, given these beliefs, player one can say "I will tell you the truth, I intend to hunt the stag, and note that if I would intend to hunt a rabbit I would tell you so." whereupon both hunt the stag. More formally, while both hunt a rabbit under language $\{\{\text{Stag, Rabbit}\}\}$, player one chooses language $\{\{\text{Stag}\}, \{\text{Rabbit}\}\}$ and intends to send message $\{\text{Stag}\}$ (but sometimes is forced to send message $\{\text{Rabbit}\}$). Both hunt the stag if message $\{\text{Stag}\}$ has been sent and hunt a rabbit if message $\{\text{Rabbit}\}$ has been sent. If player two believes player one then player one will tell the truth and not deviate from this strategy, which motivates player two to believe player one. Player two is convinced, $\{\{\text{Stag}\}, \{\text{Rabbit}\}\}$ is credible under TP. Now consider PT. Here player two will not believe player one who is saying "I will tell you the truth, I have committed to hunt the stag, and note that if I would have committed to hunt a rabbit I would tell you so.". This is because player two's best response is to copy what she believes that player one has chosen. So if player two believes player one then player one will say that she has chosen stag, even when she has chosen (or was forced to choose) rabbit. The language $\{\{\text{Stag}\}, \{\text{Rabbit}\}\}$ is not credible under PT. Consequently, player two reacts to this message of player one by choosing the action she does under the degenerate language $\{\{\text{Stag, Rabbit}\}\}$. The outcome that both hunt the rabbit can be supported. Similarly the other two NE outcomes can be supported.

The analysis of Aumann's Stag Hunt game reveals that communication helps players to coordinate on hunting the stag under TP but that it is useless, and hence unable to refine the set of Nash equilibrium outcomes under PT. This result confirms the intuition of Farrell (1988) and the findings of Charness (2000) and does not depend on which player is assigned as the interpreter (which is not true in general).

We now return to our motivating question, whether communication leads to efficiency. This is not necessarily true under PT as seen in our analysis of Aumann's Stag Hunt game informally illustrated above. Interestingly we also find that it is not true under TP as demonstrated by a 3 by 3 game of common interest. In this game the action associated to the unique efficient outcome is contained in the support of any Nash equilibrium. The message that implies that player one will not choose this action is not believable. Thus, the only credible language is the degenerate language and all three Nash equilibrium outcomes can be supported under TP. Whether or not communication leads to efficient outcomes depends on the underlying game and on whether talk is about commitments (PT) or intentions (TP). For instance, efficiency emerges in Aumann's Stag Hunt game with talk about intentions but not about commitments. On the other hand, efficiency emerges in this 3 by 3 game of common interest under talk about commitments but not about intentions. To obtain a more complete picture of the effects of communication we investigate all generic 2 by 2 games. For instance, we find that the folklore is manifested in 2 by 2 games when player one is the interpreter.

Farrell (1986, 1993) pioneered the communication literature in which messages have an intrinsic meaning. Typically communication is about private information, the stereotypical model is a sender-receiver game. In the literature on neologisms, unexpected messages are checked in terms of their credibility (self-signalling), with reasoning becoming more involved when more than one message passes this test (e.g. see Matthews et al., 1991). Baliga and Morris (2002) conduct a formal game theoretic analysis, thus avoiding plausibility checks. Notice that in “first play then talk” the subgames that start after the chosen action has been perturbed have the form of sender-receiver games. In contrast to Baliga and Morris (2002), we incorporate choice of language and allow for partial information revelation. Moreover, under “first play then talk”, private information is endogenous.

There are only few papers where communication is about intentions and messages have meaning, as we model in “first talk then play”. Farrell (1988) investigates communication about intentions in the light of rationalizability, albeit adding additional plausibility requirements and not formally defining beliefs. Lo (2007) formally analyzes elimination of weakly dominated strategies for a rich class of messages, providing intricate conditions for ruling out messages that are “opposite” to each other. She finds that a unique outcome is selected in Battle of Sexes but not in Aumann’s Stag Hunt game, the latter result being difficult to interpret. Farrell and Rabin (1996) first treat intentions as if they are private information, requiring self-signalling, and then add a condition (self-committing) that ensures that players behave according to their intentions. According to our formalization, self-signalling is not relevant for communication about intentions. Ellingsen and Östling (2010) show for the level k model that there is always more coordination on pure Nash equilibria when there is one way communication. Demichelis and Weibull (2008) consider evolution in symmetric games under two-sided communication.

Truth can be incorporated in different ways, as seen in the papers highlighted above. The two last papers assume lexicographic preferences for truth. Neologisms build on informal plausibility arguments. Baliga and Morris (2002) restrict attention to equilibria in which all information is transmitted. Other approaches include Chen (2004) who assumes that senders tell the truth with positive probability and Kartik et al. (2007) where there is a cost of telling a lie. In our paper we assume that the receiver believes that the sender tells truth, provided this is possible under the given language. Otherwise both behave as if there is a single message and truth-telling trivially holds. In contrast to Baliga and Morris (2002) this also puts discipline on out of equilibrium behavior.

There is a closely related paper by Zultan (2012), albeit where messages have no meaning, in which a game with multiple selves is proposed to account for the findings of Charness (2000). Informally it is claimed that a standard game-theoretic model will not suffice.[2] The focus is on sequential equilibria in which information is transmitted. These do not exist if the action is chosen before the message is sent, but exist if the message is sent first. Note that this does not mirror the findings of Charness (2000), even if one assumes that players select among those equilibria in which information is transmitted. This is because inefficient equilibria exist in which information is transmitted when the message is sent first.[3]

There is also experimental evidence that adding one-sided pre-play communication increases efficiency (see Cooper et al. (1989, 1992), Blume and Ortmann (2007)). An interesting connection between our model and the experiments of Weber et al. (2004) is presented in the conclusion of this paper.

In Section 2 we present the primitives of our model. Section 3 contains the definition of TP equilibrium. This is illustrated by 2 by 2 examples in Section 3.4. General results for TP equilibria for 2 by 2 games are stated

[2] Note that Baliga and Morris (2002) do not to consider the complete information setting (talk about intentions) as they find it difficult to formalize their intutions in that context (see page 467 in their paper).

[3] Let players coordinate on the mixed Nash equilibrium when message m is sent. If any other message is sent assume that they coordinate on the inefficient pure strategy Nash equilbrium.

and proven in Section 3.5. We analyze some larger games in Section 3.6. In Section 3.7, we give sufficient conditions under which communication yields efficiency in TP. Section 4 contains the definition of PT equilibrium and follows exactly the same structure as section 3. In Section 5 we conclude.

2 Preliminaries

2.1 The Underlying Game

Let Γ be a two player simultaneous move game with finite action sets S_j and von Neumann-Morgenstern utility functions $u_j : S_1 \times S_2 \to \mathbb{R}$ for player $j = 1, 2$. For a finite set X let ΔX be the set of probability distributions over X and let $C(\xi) = \{x \in X : \xi(x) > 0\}$ be the support of $\xi \in \Delta X$. $z \in \mathbb{R}^2$ is a Nash equilibrium outcome if there is a Nash equilibrium $\sigma \in \Delta S_1 \times \Delta S_2$ of Γ such that $u_j(\sigma) = z_j$ for $j = 1, 2$. z^* is the favorite Nash equilibrium outcome for player j if there is no Nash equilibrium outcome z such that $z_j > z_j^*$.

The game is called generic if $u_j(s) \neq u_j(s')$ holds for all $s, s' \in S_1 \times S_2$ and $j = 1, 2$. Note that a generic 2 by 2 game either has one Nash equilibrium or three Nash equilibria, in the latter case two are pure and one is mixed.

2.2 Communication

Communication is one-sided, from player one as sender to player two as receiver, leaving no possibility for player two to give feedback or even to respond. The language can be considered as the context in which communication takes place. This context is chosen by the interpreter who is one of the two players. i denotes the index of the player who as part of the description of the communication game has been assigned to be the interpreter.

The language defines the possible messages that player one can send.

Sending a message bears no cost. In many models of communication, messages have no meaning in which case the language would simply be a finite or infinite set, each element would be called a message. We wish to present a model in which one can investigate whether communication can be truthful and which outcomes truthful communication will yield. Given that we are considering communication before playing a game, without any future implications, telling the truth will connect messages sent to play in the game. As only player one is sending a message, we are talking about the play of player one. In particular, this could be about the action that player one wishes to choose, or a subset of actions that player one will choose from, or it could be about the particular way that player one is mixing among the different actions. However, we do not wish to model communication about mixed actions. Mixed actions are not verifiable and not a natural subject for communication. Thus we consider communication about the particular action or about a subset of actions. To communicate a subset can make sense if player one is mixing between actions and when player one does not wish to completely reveal how she is playing in the game. Consequently, messages are subsets of the set of actions of player one.

Formally, messages are elements of a partition L of S_1. This partition is called a language. Formally, L is a *language* if $L : S_1 \rightarrow\!\!\rightarrow S_1$ is a correspondence such that (i) $S_1 = \cup_{s_1 \in S_1} L(s_1)$, (ii) $L(s_1) \cap L(s_1') \neq \emptyset$ implies $L(s_1) = L(s_1')$ and (iii) $s_1 \in L(s_1)$ $\forall s_1 \in S_1$. The set of all languages is denoted by $\mathcal{L}$. Languages will be chosen by the interpreter. While we formally allow for randomizing over languages, hence choices in $\Delta\mathcal{L}$, we focus on situations in which language choices are *deterministic*, i.e. the interpreter puts all weight on a single element of $\mathcal{L}$. A message from L is a subset m of S_1 such that $m = L(s_1)$ for some $s_1 \in S_1$. To ease on notation, we identify L with its image $\{L(s_1), s_1 \in S_1\}$, thus each message m is an element of L. The degenerate language $\{S_1\}$ that contains a single element can be interpreted as there being no communication. At the opposite extreme, the language that contains only singletons, so $L(s_1) = \{s_1\}$ for all $s_1 \in S_1$, may be interpreted as complete truth-telling. These two languages will thus be referred to as “no communication” and “complete truth-telling”.

We consider two scenarios for when communication takes place. In “first talk then play” player one first sends a message to player two and then both simultaneously play Γ. In “first play then talk” player one first privately chooses an action in Γ and then sends a public message to player two after which player two chooses an action in Γ.

3 First Talk Then Play

We first model communication that occurs before either player chooses an action. First the interpreter chooses the language L. Then player one privately chooses a message m from this language L. The message $m' \in L$ actually sent to player two is possibly different as messages are perturbed. With a given probability $\varepsilon \in (0,1)$ a message from L is drawn from a given distribution η^L with full support on the set of possible messages L and sent in place of m. As η^L does not depend on the message $m \in L$ chosen, it is as if ε is the probability that player one is not allowed to choose a message. It is common knowledge which message has been sent. Finally, conditional on the chosen language and observed message both players simultaneously choose an action.

The above defines the following game, denoted by $\Gamma^{TP}(\varepsilon, \eta, i)$:

1. Player i (the interpreter) chooses a language $L \in \mathcal{L}$ and communicates it to the other player.

2. Player one privately chooses a message $m \in L$.

3. A message m' is drawn with probability $(1-\varepsilon)1_{\{m'=m\}}+\varepsilon\eta^{L}(m')$ and observed by players one and two.

4. Simultaneously, player one chooses s_1 and player two chooses s_2.

5. Payoffs are realized, where player j receives payoff $u_j(s_1,s_2)$, $j=1,2$.

We refer to m chosen in stage 2 as the *intended message*, and m' observed by both in stage 3 as the *realized message*. Perturbations are added to capture strategic considerations that would arise in more realistic communication. If the sender had the opportunity to justify the message sent, she could discuss the circumstances that would lead her to send other messages. The receiver would react to each message, the sender anticipating this, etc.. Perturbations in message sending explicitly creates scenarios in which unintended messages are sent, thereby ignoring reasons for sending alternative messages and focussing on equilibrium behavior conditional on which message has been sent. Of course, these perturbations can be explicitly interpreted as incomplete information about whether or not player one is free to choose a message. They have not been added to model misunderstandings in communication, which is a separate research topic, and do not have this interpretation as the realized message is common knowledge.

3.1 The Strategies

We now introduce the notation for the possibly mixed strategies used in $\Gamma^{TP}(\varepsilon,\eta,i)$. Let L_i be the mixed language choice of the interpreter in stage 1, so $L_i\in\Delta\mathcal{L}$. We call L_i *deterministic* if L_i puts all weight on a single language. Given language $L\in\mathcal{L}$ chosen by the interpreter in stage 1 let $m_1^L\in\Delta L$ be the mixed message sent by player one in stage 2 and let $m_1=(m_1^L)_{L\in\mathcal{L}}$. The action chosen by player one in stage 4 may depend on her intended message m in stage 2 and on the message m' realized in stage 3. Given our equilibrium concept introduced below, we do not need to consider the possibility of player one conditioning her action on her intended message. So in order to simplify notation we assume that the action chosen by player one in state 4 only depends on the realized message in stage 3 and on the language chosen in stage 1. Accordingly, let $\sigma_1^L(m')$ be the mixed action of player one in stage 4 after message $m' \in L$ has been realized in stage 3, so $\sigma_1^L : L \to \Delta S_1$. Concerning player two, let $\sigma_2^L(m')$ be the mixed action of player two in stage 4 given the language L chosen by the interpreter in stage 1 and the message m' received in stage 3, so $\sigma_2^L : L \to \Delta S_2$. We write $\sigma_j = (\sigma_j^L)_{L\in\mathcal{L}}$ for $j = 1, 2$. Hence, a strategy profile in the game $\Gamma^{TP}(\varepsilon, \eta, i)$ is a tuple $(L_i, m_1, \sigma_1, \sigma_2)$.

3.2 Credibility

We focus on equilibria in which player one truthfully communicates her intentions. We do not impose truth-telling, but only consider situations where player one is not strictly better off by not telling the truth. Whether or not the truth is told will depend on which message is sent and on why other messages are not sent. Hence it depends on the entire language. To define credibility, we do not consider what player one actually does, but whether or not there are strategies for player one such that player two can believe her. So we ignore in the following definition the incentives underlying the act of choosing a message and qualify a language as credible if each message belonging to this language can be believed by player two.

Definition 1 *We say that a language $L \in \mathcal{L}$ is* ***credible*** *if for each $m \in L$ there exists $\tau \in \Delta S_1 \times \Delta S_2$ such that $C(\tau_1) \subseteq m$ and τ is a Nash equilibrium of Γ.*

It follows from the definition that “no communication” is a credible language.

3.3 Talk then Play Equilibrium

We now present our equilibrium concept which in essence only requires in addition to subgame perfection that messages are believed whenever possible and ignored otherwise. Details will become clear latest when discussing simple examples in Section 3.4.

Definition 2 *$(L_i, m_1, \sigma_1, \sigma_2)$ is called a* ***talk then play equilibrium*** (TPE) *if*

1. *$(L_i, m_1, \sigma_1, \sigma_2)$ is a* ***subgame perfect*** *Nash equilibrium of the game $\Gamma^{TP}(\varepsilon, \eta, i)$,*

2. *L_i is credible and deterministic,*

3. *$C\left(\sigma_1^L(m')\right) \subseteq m'$ for each credible $L \in \mathcal{L}$ and each message $m' \in L$,*

4. *$\sigma_j^L = \sigma_j^{\{S_1\}}$ for $j = 1, 2$ if L is not credible.*

Given condition 1, the language L_i is chosen optimally by the interpreter in stage 1, anticipating $(m_1, \sigma_1, \sigma_2)$. Moreover, $(m_1^L, \sigma_1^L, \sigma_2^L)$ has to be a Nash equilibrium for each language L. In particular, this means that $\left(\sigma_1^L(m'), \sigma_2^L(m')\right)$ is a Nash equilibrium for each language L and for each message $m' \in L$. Conditions 2 and 3 ensure truth-telling in equilibrium. Condition 2 also requires that L_i is deterministic. This restriction comes at no loss of insight as there is no value added to choosing a mixed language in $\Gamma^{TP}(\varepsilon, \eta, i)$. Formally, if $(L_i, m_1, \sigma_1, \sigma_2)$ is a TPE and $L \in C(L_i)$ then $(L, m_1, \sigma_1, \sigma_2)$ is a TPE. Condition 3 imposes that communication is truthful whenever the language is credible, and condition 4 specifies that player two otherwise acts as if the interpreter has chosen “no communication”. Our definition of credibility ensures that condition 3 can be satisfied for each credible language.

We immediately obtain the following equivalent statement that allows us then connect to the literature.

Proposition 1 *$(L_i, m_1, \sigma_1, \sigma_2)$ is a TPE if and only if*

1. *$L_i \in \arg\max_{L \text{ is credible}} \left\{ \sum_{m' \in L} \left((1-\varepsilon)\, m_1^L(m') + \varepsilon \eta(m') \right) \cdot u_i(\sigma_1^L(m'), \sigma_2^L(m')) \right\}$,*

2. *for each credible L and each message $m' \in L$,*

   (a) $C\left(\sigma_1^L(m')\right) \subseteq m'$,

   (b) $u_2(\sigma_1^L(m'), \sigma_2^L(m')) \geq u_2(\sigma_1^L(m'), s_2)$ for all $s_2 \in S_2$,

   (c) $u_1(\sigma_1^L(m'), \sigma_2^L(m')) \geq u_1(s_1, \sigma_2^L(m'))$ for all $s_1 \in S_1$,

   (d) $\sum_{\bar{m} \in L} m_1^L(\bar{m})\, u_1(\sigma_1^L(\bar{m}), \sigma_2^L(\bar{m})) \geq u_1(s_1, \sigma_2^L(m'))$ for all $s_1 \in S_1$,[4]

3. *$\sigma_j^L = \sigma_j^{\{S_1\}}$ for $j = 1, 2$ if L is not credible.*

Note that condition 2 (c) states that once a message has been sent then there is no incentive for player one to deviate from her intentions. This is the *self-committing property* (Farrell, 1986, 1993, Baliga and Morris, 2002). Condition 2 (d) requires that player one, once anticipating later choices, does not intend to send a different message. This is different than the *self-signalling property* (Farrell, 1986, 1993, see also Baliga and Morris, 2002) as the alternative of choosing a different message is evaluated when anticipating how player two will react.

As perturbations are assumed to be small, formally ε is considered small, we describe the outcome of a TPE in terms of payoffs realized in the event that the realized and the intended message coincide.

Definition 3 *z is a TPE outcome if there exists $\bar{\varepsilon} > 0$ such that for any $\varepsilon \in (0, \bar{\varepsilon})$ there is a TPE such that $u_j(\sigma_1^{L_i}(m_1^{L_i}), \sigma_2^{L_i}(m_1^{L_i})) = z_j$ for $j = 1, 2$.*

Note that in generic games, as η only enters condition 1 in Proposition 1, the TPE outcome does not depend on η when ε is sufficiently small.

[4] It is enough to require this inequality to hold only for $s_1 = \sigma_1^L(m')$ for all $m' \in L$.

3.4 Examples

In the following we investigate several simple games. All arguments do not depend on the specific nature of the perturbations.

3.4.1 Stag Hunt (TP)

Consider the version of the Stag Hunt game as discussed by Aumann (1990) shown in Figure 1.

	S	R
S	$9,9$	$0,8$
R	$8,0$	$7,7$

Figure 1: Aumann's Stag Hunt game

The following analysis does not depend on who is the interpreter.

(i) "Complete truth-telling" is credible. If player two believes that player one will truthfully reveal the pure action she intends to choose then player one will tell the truth. I.e., $\{\{S\},\{R\}\}$ is a credible language. "No communication" $\{\{S,R\}\}$ is also credible, by definition.

(ii) Under "complete truth-telling", both play S whenever $\{S\}$ is sent, and both play R whenever $\{R\}$ is sent. Thus, player one intends to send $\{S\}$ under this language, which is beneficial for both players. But as there can be perturbations in message sending, "complete truth-telling" is only best if "no communication" is followed by (R,R) or by the mixed Nash equilibrium of the underlying game.

(iii) Thus, there are two TPE in which the interpreter chooses complete truth-telling and where choice of "no communication" does not result in (S,S) . There is a third TPE in which the interpreter chooses "no communication" which is followed by beliefs that (S,S) will be chosen.

To summarize, we obtain that communication leads to efficiency in this game.

3.4.2 Hawk Dove (TP)

Consider now the Hawk Dove game (or Game of Chicken) shown in Figure 2 which has three Nash equilibrium outcomes which are all efficient.

	H	D
H	$-1, -1$	$2, 0$
D	$0, 2$	$1, 1$

Figure 2: Hawk Dove game

(i) Both $\{\{H, D\}\}$ and $\{\{H\}, \{D\}\}$ are credible, the latter follows from the fact that (H, D) and (D, H) are Nash equilibria of the underlying game. If $\{\{H\}, \{D\}\}$ is chosen then player one will intend to send $\{H\}$ and induce play of (H, D) most of the time.

(ii) Assume that player one is the interpreter. She will choose language $\{\{H, D\}\}$ if this is followed by beliefs that (H, D) will be played. Otherwise she will choose $\{\{H\}, \{D\}\}$. So we find two TPE very similar to those found in the Stag Hunt Game. One involves "no communication" as players believe that (H, D) will be played. The other one involves complete truth-telling as player one "fears" that "no communication" is followed by either play of (D, H) or of the mixed Nash equilibrium.

(iii) Now assume that player two is the interpreter. Then there are three TPE, one associated to each Nash equilibrium of the underlying game. No communication will arise in equilibrium if it is followed by (D, H) or the mixed Nash equilibrium. Complete truth-telling will arise, leading most of the time to outcome (H, D), if no communication is followed by play of (H, D). Note in this case that player two as interpreter strictly prefers "complete truth-telling" as perturbations in message sending sometimes lead to her preferred outcome (D, H) . This is an instance where the explicit model of perturbations yields predictions and insights.

We summarize. Player one obtains her favorite outcome if she is the interpreter. If instead player two is the interpreter then the addition of communication does not reduce the set of equilibrium outcomes. This example shows how the power of the sender can be weakened if the receiver is the interpreter. It is also an example for how communication can fail to help coordinate beliefs when player two is the interpreter.

3.4.3 Battle of Sexes (TP)

Consider now Battle of Sexes as shown in Figure 3.

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

Figure 3: Battle of Sexes

(i) If player one is the interpreter we find the analogous outcomes as in the Stag Hunt and Hawk Dove games. The possibility to communicate leads to the most favorable outcome for player one. The equilibrium language can involve either "complete truth-telling" or "no communication".

(iii) Assume that player two is the interpreter. $(3, 1)$ is the only TPE outcome that involves complete truth-telling, supported either by (T, L) or the mixed Nash equilibrium when there is no communication. There is a TPE with no communication which is followed by (B, R) . However, unlike in the Hawk Dove game, the mixed equilibrium is not a TPE outcome in Battle of Sexes. If no communication is followed by the mixed Nash equilibrium then player two strictly prefers to choose complete truth-telling as she strictly prefers outcome $(3,1)$ to the mixed Nash equilibrium outcome.

To summarize, communication selects one of the two pure Nash equilibria. Player one gets her favorite outcome if she is the interpreter. Moreover, as in all the previous examples, TPE outcomes are efficient.

3.4.4 A Simple 2 by 2 Game (TP)

The game in Figure 4 is used to show that TPE outcomes need not be efficient when player two is the interpreter.

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

Figure 4: A Simple 2 by 2 Game

This game has the three Nash equilibrium outcomes $(3,1)$, $(2,3)$ and $\left(\frac{3}{2},\frac{3}{2}\right)$. Assume that player two is the interpreter. Then there is a TPE in which player two chooses "no communication" where this leads to outcome $\left(\frac{3}{2},\frac{3}{2}\right)$. An inefficient Nash equilibrium outcome can arise because it is strictly preferred by player two to the favorite Nash equilibrium of player one. At the same time, both $(3,1)$ and $(2,3)$ are TPE outcomes which are efficient.[5]

3.5 General Results for 2 by 2 games (TP)

In this section we investigate whether our intuition gathered in the above examples holds more generally in generic 2 by 2 games. The first result in fact holds also for larger games.

[5] It is similarly easy to construct a 3 by 3 game, in which all Nash equilibria are in pure strategies, that has an ineffcient TPE outcome when player two is the interpreter.

Proposition 2 (existence) *In any game there exists a TPE in which in equilibrium "no communication" is followed by play of the Nash equilibrium that is most favorable to the interpreter.*

Proof. If "no communication" is followed by play of the Nash equilibrium in which the interpreter is best off then there is no incentive for the interpreter to choose a different language. ■

One might interpret this result as evidence for the power of the interpreter. However there might be other TPE outcomes. In fact, as we have seen in Figure 4, if player two is the interpreter, then there is also a TPE in which player two gets her worst equilibrium payoff . In Proposition 5 below we investigate the power that player one has as interpreter. However first we consider two other properties of TPE outcomes.

Proposition 3 (Nash) *In any generic* 2 *by* 2 *game any TPE outcome is a Nash equilibrium outcome.*

Proof. If the interpreter chooses "no communication" then the outcome must be a Nash equilibrium by conditions 2 (b) (c) in Proposition 1. If the interpreter chooses "complete truth-telling" then by genericity player one strictly prefers one of the messages over the other. When ε is small, most of the time the preferred intended message will also be the realized message. By conditions 2 (b) (c) of Proposition 1 it follows that the TPE outcome is a Nash equilibrium outcome. ■

We hasten to point out, using the game in Figure 5, that TPE outcomes need not be Nash equilibrium outcomes. In this game there is a TPE in which the interpreter chooses "complete truth-telling", then player one intends to send $\{T\}$ and $\{B\}$ equally likely. The resulting TPE outcome $(1, 3/2)$ is not a Nash equilibrium outcome of Γ.[6]

[6] It is however a correlated equilibrium outcome of Γ.

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

Figure 5: A non generic 2 by 2 Game

Proposition 4 (efficiency) *Given any generic* 2 *by* 2 *game, if player one is the interpreter then any TPE outcome is efficient within the set of Nash equilibrium outcomes.*

The proof of Proposition 4 is a simple consequence of the next proposition.

Proposition 5 (power) *Given any generic* 2 *by* 2 *game, if player one is the interpreter then any TPE outcome is the favorite Nash equilibrium outcome of player one.*

Proof. If there is only one Nash equilibrium, then the statement is trivial by Proposition 3. Assume there are three Nash equilibria. Thus both "no communication" and "complete truth-telling" are credible. As the game is a 2 by 2 game it is easy to see that the favorite Nash equilibrium outcome of player one is supported by a pure strategy Nash equilibrium. Consequently, provided ε is sufficiently small, player one as interpreter will choose "complete truth-telling", yielding with high probability the favorite Nash equilibrium outcome of player one. As TPE outcomes only refer to sufficiently small ε the proof is completed. ■

3.6 Examples involving Larger Games

Next we move to larger games to discover new aspects of credible communication. We first present two examples that show that Propositions 4 and 5 do not generalize to all larger generic games.

3.6.1 An Augmented Hawk Dove Game (TP)

Consider the augmented Hawk Dove game shown in Figure 6 where player two has an additional action R. The Nash equilibrium outcomes in this game are $\left(\frac{1}{4},3\right)$, $(0,2)$ and $\left(\frac{1}{2},\frac{1}{2}\right)$.

	H	D	R
H	$-1,-1$	$2,0$	$\frac{1}{4},3$
D	$0,2$	$1,1$	$-2,-3$

Figure 6: An Augmented Hawk Dove Game

(i) Assume that player one is the interpreter. There is a TPE in which player one chooses "no communication", anticipating play of $\frac{1}{2}[H]+\frac{1}{2}[D]$ by both, leading to outcome $\left(\frac{1}{2},\frac{1}{2}\right)$. There is a TPE with "complete truth-telling" leading to outcome $\left(\frac{1}{4},3\right)$, where "no communication" is followed by play of (D,H). There are no other TPE outcomes. In particular, the favorite outcome of player one is not the only TPE outcome. This shows that Proposition 5 does not generalize to larger generic games.

(ii) If instead player two is the interpreter then there is a unique TPE outcome, namely $\left(\frac{1}{4},3\right)$. It can be supported by "complete truth-telling" if "no communication" is followed by either of the two other Nash equilibrium outcomes. It can similarly be supported by "no communication" that is followed by play of (H,R).

Note that in all previous examples, player two as interpreter could not guarantee her favorite outcome unless it was also the favorite outcome of player one. The reason why player two as interpreter can ensure her favorite outcome in this game is that it will be chosen by player one under "complete truth-telling".

3.6.2 A 3 by 3 Common Interest Game (TP)

The game shown in Figure 7 demonstrates how communication can be useless even if the game has common interests.

	L	N	R
T	$5,5$	$0,0$	$-3,-3$
M	$-1,-1$	$1,1$	$2,2$
B	$4,4$	$-2,-2$	$3,3$

Figure 7: A Common Interest Game

(T,L) is a pure strategy Nash equilibrium that leads to the unique efficient outcome.[7] It is natural that player one wants to say “I will play T”. However, each of the other two Nash equilibria of this game have T in the support of the corresponding equilibrium strategy of player one.[8] This means that player one cannot truthfully communicate that she will not be playing T. Consequently, only $\{\{T,M,B\}\}$ is a credible language. Regardless of who is the interpreter, nontrivial information about intentions cannot be transmitted under credible communication in this game.[9] In particular this shows that Proposition 4 does not extend to larger generic games.

[7] In fact, T is self-committing and the game satisfies self-signalling (Farrell, 1986, 1993).

[8] The other two mixed Nash equilibria τ and ρ are given by

$$\tau_1(T)=2/7, \tau_1(M)=5/7, \tau_1(B)=0, \tau_2(L)=1/7, \tau_2(N)=6/7, \tau_2(R)=0$$

and

$$\rho_1(T)=4/15, \rho_1(M)=43/60, \rho_1(B)=1/60, \rho_2(L)=4/15, \rho_2(N)=31/60, \rho_2(R)=13/60$$

with corresponding outcomes $5/7$ and $41/60$.

[9] We hasten to point out that if one enriches the set of messages and allows for player one to communicate which mixed action she will be choosing then one would obtain efficiency in any common interest game.

3.7 Efficiency and Communication (TP)

We refrain from a general analysis of all larger games. Instead we return briefly to our initial motivation, efficiency. As we observed in the game shown in Figure 7, credible communication can only lead to efficient outcomes if one can credibly talk about counterfactuals. Moreover, following the insights from the analysis of the game in Figure 4 we can only expect a general result on efficiency if player one is the interpreter.

Proposition 6 (efficiency) *Let ξ be the Nash equilibrium associated to the favorite Nash equilibrium outcome z of player one. Assume that (i) there is no other Nash equilibrium of the game in which player one only chooses actions belonging to $C(\xi_1)$, and (ii) there is some Nash equilibrium ξ' of Γ such that $C(\xi_1) \cap C(\xi'_1) = \emptyset$. If player one is the interpreter then z is the unique TPE outcome.*

Proof. Condition (ii) implies that $\{C(\xi_1), S_1 \backslash C(\xi_1)\}$ is a credible language. Condition (i) implies that ξ is played whenever message $C(\xi_1)$ is sent. Consequently, player one can ensure outcome z with high probability by choosing language $\{C(\xi_1), S_1 \backslash C(\xi_1)\}$. The proof then follows from the fact that z is the favorite Nash equilibrium outcome of player one. ■

4 First Play then Talk

We now consider the situation where communication takes place after player one but before player two chooses an action. First player one chooses her action, which is not observable by player two. This action is then perturbed. With a given probability $\varepsilon \in (0,1)$ the action chosen by player one is replaced with one drawn from the distribution $\eta \in \Delta S_1$ with $C(\eta) = S_1$. After player one's action is realized (possibly an unintended action), we will think of this action as player one's realized *type*, which is her private information, thus unknown to player two. The two players then meet to communicate. The interpreter chooses the language, thereafter player one chooses a message from this language. Finally player two chooses her action.

The above defines a game $\Gamma^{PT}(\varepsilon, \eta, i)$:

1. Player one privately chooses an action $s_1 \in S_1$.
2. An action $s_1' \in S_1$ is realized with probability $(1-\varepsilon)1_{\{s_1'=s_1\}} + \varepsilon\eta(s_1')$, only player one observes this realization.
3. Player i (as the interpreter) publicly chooses a language $L \in \mathcal{L}$.
4. Player one sends a message $m \in L$ to player two.
5. Player two chooses an action $s_2 \in S_2$.
6. Payoffs are realized, where player j receives payoff $u_j(s_1', s_2)$, $j = 1, 2$.

Once again perturbations are added to capture strategic considerations that would arise in more realistic communication scenario. The sender would wish to talk about the information contained in her message and what she would have said if she had chosen a different action. Perturbations explicitly introduce events in which the sender has chosen a different action, and allow the receiver to anticipate how the sender behaves in these alternative scenarios, and vice versa. Of course perturbations can explicitly be interpreted as incomplete information about whether or not player one is free to choose an action.

4.1 The Strategies

Let $\sigma_1 \in \Delta S_1$ be the mixed action of player one in stage 1. Consider stage 3 and assume that player one is the interpreter. Then her choice of the language may depend on her intended action in stage 1 and on the realized action in stage 2. Note however that her intended choice has no payoff relevance and is not observable to player two. Hence there is no value added to conditioning on it. So in order to simplify notation we assume in this case, where player one is the interpreter, that the language choice only depends on the realized action in stage 2. So let $L_1\left(s_1'\right)$ be the mixed language chosen in stage 3 after action s_1' has been realized in stage 2, $L_1 : S_1 \to \Delta\mathcal{L}$. We now define player two's belief $\nu_1^L \in \Delta S_1$ about the realized action s_1' given that player one has chosen language L. Below we will focus on language choices of player one that do not depend on her realized type. In view of this, we will assume that player two treats the language choices of player one as being independent of the realized type. Hence, we set ν_1^L equal to the ex-ante beliefs $(1-\varepsilon)\sigma_1+\varepsilon\eta$. Note that these will also be the beliefs of player two, denoted by ν_2^L, if instead player two is the interpreter.

In stage 4, player one chooses a mixed message m_1^L belonging to the language L chosen in stage 3 given that action s_1' is realized in stage 2, so $m_1^L : S_1 \to \Delta L$ and $m_1 = (m_1^L)_{L\in\mathcal{L}}$. Here we rule out, consistent with our approach above, that player one conditions her message on what happened in stage 1.

In stage 5, player two chooses a mixed action $\sigma_2^L\left(m\right)$ that depends on the language L chosen in stage 3 and on the message m received in stage 4, so $\sigma_2^L : L \to \Delta S_2$ and $\sigma_2 = (\sigma_2^L)_{L\in\mathcal{L}}$.

Hence a strategy profile in the game $\Gamma^{PT}\left(\varepsilon,\eta,i\right)$ is described by $(\sigma_1, L_i, m_1, \sigma_2)$.

4.2 Credibility

It is useful to view communication in stage 4 and the consequent choice of player two in stage 5 as a sender-receiver game under incomplete information. Consider some language L, that has been chosen in stage 3, and some beliefs

$\nu \in \Delta S_1$ of player two about the realized type of player one as determined in stage 2. Player one, who knows her realized type, sends a message from this language to player two who then makes a choice. Denote this (auxiliary) sender-receiver game by $\Gamma(L, \nu)$. Specifically, $\Gamma(L, \nu)$ is defined as follows:

i. Player one's action s_1' is chosen by *nature* according to ν and revealed only to player one,

ii. player one sends a messages from L to player two,

iii. player two chooses an action s_2 from S_2,

iv. payoffs are realized, where player j receives payoff $u_j(s_1', s_2)$ for $j = 1, 2$.

A strategy for player one is given by $\tau_1 : S_1 \to \Delta L$, a strategy for player two is given by $\tau_2 : L \to \Delta S_2$.

We now define credible languages. As in TP, a language is credible if each message belonging to this language can be believed by player two.

Definition 4 *$L \in \mathcal{L}$ is called a **credible language** given $\nu \in \Delta S_1$ if there exists a Nash equilibrium τ of the game $\Gamma(L, \nu)$ in which player one tells the truth, so where $\tau_1(s_1) = L(s_1)$ for all $s_1 \in S_1$.*

Trivially, "no communication" is a credible language.

4.3 Play then Talk Equilibrium

We now present our equilibrium concept which requires that communication is truthful and that languages convey no information about the type of the sender.

Definition 5 *$(\sigma_1, L_i, m_1, \sigma_2)$ is called a **play then talk equilibrium*** (PTE) *if*

1. *$(\sigma_1, L_i, m_1, \sigma_2)$ is a Nash equilibrium of the game $\Gamma^{PT}(\varepsilon, \eta, i)$,*

2. *L_i is credible given $(1-\varepsilon)\sigma_1 + \varepsilon\eta$, deterministic and if $i = 1$ then L_1 is constant,*

3. *for each $L \in \mathcal{L}$ that is credible given $(1-\varepsilon)\sigma_1 + \varepsilon\eta$,*

   (a) $m_1^L(s_1') = L(s_1')$ holds for all $s_1' \in S_1$,

   (b) (m_1^L, σ_2^L) is a Nash equilibrium of $\Gamma(L, (1-\varepsilon)\sigma_1 + \varepsilon\eta)$,

4. *$\sigma_j^L = \sigma_j^{\{S_1\}}$ for $j = 1, 2$ if L is not credible given $(1-\varepsilon)\sigma_1 + \varepsilon\eta$.*

Condition 1 identifies that we are interested in Nash equilibria.[10] The last statement in condition 2 ensures when player one is the interpreter that the equilibrium language L_1 does not contain any information about what happened in stage 2. Given this restriction there is no loss of generality to assume that the equilibrium language choice is deterministic as required in condition 2. Conditions 2 and 3 (a) ensure that there is truth-telling in equilibrium. Furthermore, condition 3 (a) ensures that there is truth-telling whenever the chosen language is credible. Condition 4 specifies that both players act as if the interpreter has chosen "no communication" whenever the chosen language is not credible. Conditions 3 (b) and 4 ensure that players best respond to each other. Beliefs of player two are set equal to $(1-\varepsilon)\sigma_1 + \varepsilon\eta$ on the equilibrium path (condition 2) and off the equilibrium path (conditions 3 and 4). Clearly, when player two is the interpreter then beliefs are given by $(1-\varepsilon)\sigma_1 + \varepsilon\eta$ independently of which language has been chosen. Now assume that player one is the interpreter. As L_1 is required to be constant (condition 2) these are also the beliefs under the equilibrium language L_1. We also maintain these beliefs whenever player one has chosen a language $L \neq L_1$. It is as if player two assumes when player one chooses a

[10] Unlike in TP, subgame perfection does not select among the Nash equilibria.

different language, that this choice is not conditional on the realized action. With these beliefs, languages chosen out of equilibrium also do not contain any information for player two about actions chosen or realized. Of course, player one as interpreter is allowed to deviate and choose different languages for different realized actions.

We show some properties of PTE.

Proposition 7 *Any PTE is a Perfect Bayesian equilibrium of* $\Gamma^{PT}(\varepsilon, \eta, i)$.

Proof. Player two formulates her beliefs for any credible language by updating her prior $(1-\varepsilon)\sigma_1 + \varepsilon\eta$ when receiving a message. When the interpreter chooses a non credible language, player two keeps her prior belief $(1-\varepsilon)\sigma_1 + \varepsilon\eta$ regardless of which message she receives. The conditions of our definition ensure that player two best responds to these beliefs. ■

Given this result we did not explicitly require in our definition for a PTE to be a Perfect Bayesian equilibrium.

We present an equivalent formulation of the PTE definition to connect better to the literature. As the equilibrium language used in a PTE is constant, the updated beliefs of player two, denoted by $p_2(s_1|m, L)$, do not depend on the language and are hence given by

$$p_2(s_1|m) = \frac{(1-\varepsilon)\,\sigma_1(s_1) + \varepsilon\eta(s_1)}{\sum_{s_1' \in m}\left((1-\varepsilon)\,\sigma_1(s_1') + \varepsilon\eta(s_1')\right)}.$$

Let us denote by $B_2(m) = \arg\max_{s_2 \in S_2} \sum_{s_1 \in S_1} p_2(s_1|m) u_2(s_1, s_2)\}$ the set of best responses of player two to message $m \subseteq S_1$ given her updated beliefs $p_2(s_1|m)$ about s_1.

Proposition 8 $(\sigma_1, L_i, m_1, \sigma_2)$ *is a PTE if and only if there exists* $\overline{L} \in \mathcal{L}$ *such that*

1. $C(\sigma_1) \subseteq argmax_{s_1 \in S_1} u_1(s_1, \sigma_2^{\overline{L}}(\overline{L}(s_1)))$,

2. (a) *if* $i = 1$ *then* $L_1(s_1) = \overline{L}$ *and* $\bar{L}$ *solves* $\max_{L\in\mathcal{L}} u_1(s_1, \sigma_2^L(L(s_1)))$ *for all* $s_1 \in S_1$,

   (b) *if* $i = 2$ *then* $L_2 = \overline{L}$ *and* $\bar{L}$ *solves*

$$\max_{L\in\mathcal{L}} \left\{ (1-\varepsilon) \sum_{s_1\in C(\sigma_1)} \sigma_1(s_1)\, u_2(s_1, \sigma_2^L(L(s_1))) + \varepsilon \sum_{s_1\in S_1} \eta(s_1) u_2(s_1, \sigma_2^L(L(s_1))) \right\},$$

3. *for each* $L \in \mathcal{L}$ *that is credible given* $(1-\varepsilon)\sigma_1 + \varepsilon\eta$,

   (a) $m_1^L(s_1) = L(s_1)$ *for all* $s_1 \in S_1$

   (b) $C(\sigma_2^L(m)) \subseteq B_2(m)$ *for all* $m \in L$,

   (c) *for all* $s_1 \in S_1$ *and* $m \in L$,

$$u_1(s_1, \sigma_2^L(L(s_1))) \geq u_1(s_1, \sigma_2^L(m)), \tag{1}$$

4. $\sigma_j^L = \sigma_j^{\{S_1\}}$ *for* $j = 1, 2$ *if* L *is not credible given* $(1-\varepsilon)\sigma_1 + \varepsilon\eta$.

Condition 3 (c) reflects the *self-signalling property* (Farrell, 1986, 1993). Player one wishes that player two believes that she is telling the truth and does not prefer to tell a lie and hence induce player two to choose a different best response. Note that self-signalling is formalized by Baliga and Morris (2002) (see their Definition 3) by replacing $\sigma_2^L(m)$ by s_2 in (1) and requiring this inequality to hold for all $s_2 \in S_2$. This is too strong and misleading. Player one cannot make player two choose an arbitrary action (see also our Section 4.7).

Definition 6 *z is a PTE outcome if there exists* $\overline{\varepsilon} > 0$ *such that for any* $\varepsilon \in (0, \bar{\varepsilon})$ *there is a PTE such that* $u_j(\sigma_1, \sigma_2^{L_i}(m_1^{L_i})) = z_j$ *for* $j = 1, 2$.

4.4 Examples

Next we investigate several simple games. All arguments do not depend on the specific probabilities in the perturbations.

4.4.1 Stag Hunt (PT)

Consider Aumann's Stag Hunt game in Figure 1.

(i) Only "no communication" is credible. For "complete truth-telling" to be credible player two has to believe that the realized action is S when hearing $\{S\}$. Hence, player one, who either intentionally or accidently chooses R, is better off sending $\{S\}$ than sending $\{R\}$ if she believes that player two believes her.

(ii) As only "no communication" is credible, any of the three Nash equilibrium outcomes of Aumann's Stag Hunt game can arise in a PTE.

4.4.2 Hawk Dove (PT)

We obtain similar findings for the Hawk Dove game (see Figure 2). Complete truth-telling is not credible as player one always wants player two to believe that she has chosen H. Consequently, only "no communication" is credible and any of the three Nash equilibrium outcomes can arise in a PTE.

4.4.3 Battle of Sexes (PT)

Consider now Battle of Sexes (Figure 3).

(i) Both $\{\{T\},\{B\}\}$ and $\{\{T,B\}\}$ are credible.

(ii) Regardless of which action player one has realized, both players are best off if player two learns which action this is. Hence each player, when she is the interpreter, will choose $\{\{T\},\{B\}\}$. Anticipating this, player one chooses T. This shows that only (T,L) and "complete truth-telling" arise in a PTE.

4.5 General results for 2 by 2 games (PT)

We now investigate general 2 by 2 games such as the one in Figure 8.

	L	R
T	a, e	b, f
B	c, g	d, h

Figure 8: A general 2 by 2 Game

Proposition 9 (existence) *In any generic* 2 *by* 2 *game there exists a PTE.*

Proof. If only "no communication" is credible then a PTE trivially exists. So consider the case where both "no communication" and "complete truth-telling" are credible. This means that Γ has three Nash equilibria. Suppose w.l.o.g. that the two pure strategy Nash equilibria are given by (T, L) and (B, R), where $a > d$. In order for "complete truth-telling" to be credible we also need that $a > b$ and $d > c$. This then implies that both players prefer that player one truthfully reveals the realized action. Hence there is a PTE in which the interpreter chooses "complete truth-telling" and player one intends to choose T. ■

The following two propositions follow directly from the proof of Proposition 9.

Proposition 10 *In any generic* 2 *by* 2 *game, regardless of who is the interpreter, any PTE outcome is a Nash equilibrium outcome of the underlying game* Γ*.*

Proposition 11 *Consider a generic* 2 *by* 2 *game in which complete truth-telling is credible. Regardless of who is the interpreter, any PTE outcome is a subgame perfect equilibrium outcome of the perfect information game in which first player one chooses a pure action and then player two chooses a pure action.*[11]

[11] One may wish to call this the Stackelberg outcome, see also Baliga and Morris (2002).

4.6 Examples of Larger Games

4.6.1 The Augmented Hawk Dove Game (PT)

Consider the augmented Hawk Dove game in Figure 6.

(i) Both $\{\{H\},\{D\}\}$ and $\{\{H,D\}\}$ are credible languages.

(ii) Assume that player one is the interpreter. Then there are two possible PTE outcomes. There is a PTE that supports the favorite outcome $\left(\frac{1}{2},\frac{1}{2}\right)$ of player one. Player one anticipates the perturbation and mixes between H and D in a way that the probability that H is realized is equal to $\frac{1}{2}$. She then chooses $\{\{H,D\}\}$ regardless of which action is realized whereupon player two mixes equally between H and D. If instead player one chooses $\{\{H\},\{D\}\}$ then player one tells the truth. There is also a PTE that supports the favorite outcome $\left(\frac{1}{4},3\right)$ of player two even though player one is the interpreter. Player one chooses H and then, regardless of which action has been realized, chooses $\{\{H\},\{D\}\}$. This leads to outcome $\left(\frac{1}{4},3\right)$. If instead player one chooses $\{\{H,D\}\}$ then player two chooses R. Note that $(0,2)$ cannot be a PTE outcome as player one can always ensure $\frac{1}{4}$ by choosing H and then choosing $\{\{H\},\{D\}\}$.

(iii) Assume instead that player two is the interpreter. Player two is best off if she learns which action has been realized by player one. Hence she chooses $\{\{H\},\{D\}\}$. Player one anticipates this and chooses H which leads to $\left(\frac{1}{4},3\right)$. This is the unique PTE outcome.

In particular, we learn from part (ii) that Proposition 11 does not extend to larger games when player one is the interpreter (see also Proposition 12 below).

4.6.2 A Common Interest Game (PT)

Consider the game of common interest in Figure 7. Both players are strictly better off if player two knows what player one realized. Hence we obtain a unique PTE, it involves complete truth-telling and yields an efficient PTE outcome.

4.7 Efficiency and Communication (PT)

We provide a sufficient condition for efficiency that is related to the definition of self-signalling in Baliga and Morris (2002). Given our discussion of Proposition 8 in Section 4.3 we give it a different name.

Definition 7 Γ *is self-choosing (for player one) if* $u_1(s_1, s_2^*(s_1)) \geq u_1(s_1, s_2)$ *for all* $s_1 \in S_1$*,* $s_2 \in S_2$ *and* $s_2^*(s_1) \in \arg\max_{s_2' \in S_2} u_2(s_1, s_2')$ *.*

When a game is self-choosing then player one is best off, when choosing any of her actions, if player two best responds. One might also say that there is common interest in the best response behavior of player two. Note that complete truth-telling is credible whenever Γ is self-choosing. For example, common interest games are self-choosing, but the augmented Hawk Dove game is not self-choosing as $u_1(D, b_2(D)) = 0 < u_1(D, D) = 1$.

Proposition 12 *Consider a generic game that is self-choosing. Then there is a unique PTE outcome. It is the favorite Nash equilibrium outcome of player one which is efficient within the set of Nash equilibrium outcomes.*[12]

Proof. Take any PTE. Clearly, player two is weakly best off when complete truth-telling is chosen. The self-choosing property ensures that player one is also weakly best off when complete truth-telling is chosen, regardless of the realized action. This is because self-choosing implies player one has no incentives to hide some information and induce player two to play some action different from her pure best responses. So regardless of who is the

[12] The proposition remains true if one replaces genericity by the following alternative condition: $u_1(s_1, s_2^*(s_1)) = u_1(s_1', s_2^*(s_1'))$ implies $u_2(s_1, s_2^*(s_1)) = u_2(s_1', s_2^*(s_1'))$.

interpreter, there is a possibly different PTE with the same payoffs for each player where the interpreter chooses complete truth-telling. Player one then chooses the action s_1^* which maximizes $u_1(s_1, b_2(s_1))$ over all $s_1 \in S_1$. Hence, $(s_1^*, b_2(s_1^*))$ is the unique PTE outcome. In particular, the above shows that a PTE with complete truth-telling exists.

Note that $(s_1^*, b_2(s_1^*))$ is a pure strategy Nash equilibrium of Γ as $u_1(s_1^*, b_2(s_1^*)) \geq u_1(s_1, b_2(s_1)) \geq u_1(s_1, b_2(s_1^*))$. The second inequality follows as complete truth-telling is credible. So $(s_1^*, b_2(s_1^*))$ is the favorite pure strategy Nash equilibrium of player one. We now show that $(s_1^*, b_2(s_1^*))$ is the favorite Nash equilibrium outcome of player one. Here we use the self-choosing property. Assume by contradiction that there is a mixed Nash equilibrium $(\tau_1, \tau_2) \in \Delta S_1 \times \Delta S_2$ such that $u_1(\tau_1, \tau_2) > u_1(s_1^*, b_2(s_1^*))$. Then there exists $(s_1', s_2') \in S_1 \times S_2$ such that $u_1(s_1', s_2') > u_1(s_1^*, b_2(s_1^*))$. But $u_1(s_1', b_2(s_1')) \geq u_1(s_1', s_2')$ by the self-choosing property and hence $u_1(s_1', b_2(s_1')) > u_1(s_1^*, b_2(s_1^*))$ which contradicts the fact that s_1^* maximizes $u_1(s_1, b_2(s_1))$ over all $s_1 \in S_1$.

Finally, note that genericity ensures that any favorite Nash equilibrium outcome of player one is also efficient within the set of Nash equilibrium outcomes. ■

Self-choosing along with other conditions are sufficient in Baliga and Morris (2002) for the existence of an outcome that is efficient within the set of Nash equilibrium outcomes. Note however that, while we have a unique equilibrium outcome, Baliga and Morris (2002) may also have inefficient equilibria.

Efficiency cannot be guaranteed without self-choosing. To see this, replace $(1/4, 3)$ by $(1/4, 1/4)$ in the augmented Hawk Dove game, an outcome that is dominated by the mixed Nash equilibrium outcome $(1/2, 1/2)$. It is easily checked that $(1/4, 1/4)$ is a possible PTE outcome when player one is the interpreter, and the unique PTE outcome when player two is the interpreter.

5 Conclusion

Interestingly, despite the large literature on communication in games, we seem to be the first to use an equilibrium analysis to investigate the impact of truthful communication under pre-play communication (as modelled in our "first talk then play" scenario). Truthful does not mean that players are forced to tell the truth. It means that the sender is able to convince the receiver whenever she can be believed. We call this *credible communication.* Perturbations take the role of talking about counter factual evidence. Our findings show that efficiency is not guaranteed in common interest games that have more than two strategies per player. The debate raised by Aumann also necessitates that we present a model in which communication occurs during play, called "first play then talk". This model has its own value as it is the first step to understanding communication while playing sequential games of imperfect information. Results in the two models are very different and are useful to highlight how communication influences outcomes. They are both very tractable when analyzing specific games and can help understand in applications which equilibria have good properties. After all, parties will typically communicate and this should be considered formally when making predictions, instead of using it only as a motivation like in the literature on renegotiation.

Clearly communication as modelled in this paper is very specific. Once our modelling approach is well received we believe it to be important to tackle various extensions. Note that we have already explicitly considered sender receiver games with incomplete information within our model of "first play then talk". To be able to also deal with this popular class of games was an important concern when setting up our model. We find it valuable, thereby contrasting the modelling of Baliga and Morris (2002), to allow for general messages and to identify all equilibria with truth-telling, and not just those where all information is transmitted. In other words, we wish to predict outcomes in games, not to understand when all information can be transmitted. Other extensions that are easy to implement include considering the case where player two is uncertain about whether or not player one has already committed to an action and considering an n player game where only player one communicates to the others. Extensions that require more thought in terms of making the right modelling choice include two-sided communication.

Finally, note that there may be a connection to the experiments of Weber et al. (2004). In these experiments there is some evidence that the first mover in a sequential game of imperfect information is better off than in the associated simultaneous move game even if the first move is not observable by the other players. In their implementation of the Ultimatum game, the first mover gets her favorite outcome under PT (following Proposition 12 and footnote 12). This is consistent with a postulate that subjects act in the sequential version of their experiment as if there is communication between making the choices. Now consider communication prior to their treatment with simultaneous choices, as modelled in TP. Then player one gets her favorite outcome when she is the interpreter (following Proposition 6). However, when player two is the interpreter then there are many different TPE outcomes, in particular player two may receive her favorite outcome (following Proposition 2). Additional experiments would be needed in order to shed more light on this possible connection.

References

[1] R.J. Aumann, (1990), “Nash-Equilibria are not Self-Enforcing”, *in Economic Decision Making: Games, Econometrics and Optimisation* (J. Gabszewicz, J.-F. Richard, and L. Wolsey, Eds.), Amsterdam, Elsevier 201-206.

[2] S. Baliga and S. Morris (2002), “Co-ordination, Spillovers, and Cheap Talk”, *Journal of Economic Theory* **105**, 450–468.

[3] A. Blume and A. Ortmann (2007), “The effects of costless pre-play communication: Experimental evidence from games with Pareto-ranked equilibria”, Journal of Economic Theory **132**, 274–290.

[4] G. Charness, (2000), “Self-Serving Cheap Talk: A Test of Aumann’s Conjecture”, *Economic Theory* **33**, 177-194.

[5] Y. Chen (2004), “Perturbed Communication Games with Honest Senders and Naive Receivers", Journal of Economic Theory **146**, 401–424.

[6] K. Clark, S. Kay and M. Sefton (2001) “When Are Nash Equilibria Self-Enforcing? An Experimental Analysis”, *International Journal of Game Theory* **29**, 495-515.

[7] R. Cooper, D.V. DeJong, R. Forsythe and T.W. Ross (1989), “Communication in the Battle of the Sexes Game: Some Experimental Results”, *The RAND Journal of Economics* **20**, 568-587.

[8] R. Cooper, D.V. DeJong, R. Forsythe and T.W. Ross (1992), “Communication in Coordination Games”, *The Quarterly Journal of Economics* **107**. 739-771.

[9] S. Demichelis and J.W. Weibull (2008), “Language, Meaning, and Games: A Model of Communication, Coordination, and Evolution”, *American Economic Review* **98**, 1292–1311.

[10] T. Ellingsen and R. Östling (2010), “When Does Communication Improve Coordination?” *American Economic Review* **100**, 1695–1724.

[11] J. Farrell (1986), “Meaning and Credibility in Cheap Talk Games,” University of California, Berkeley, Department of Economics working paper 8609.

[12] J. Farrell (1988), “Communication, Coordination, and Nash Equilibrium”, *Economic Letters* **27**, 209-214.

[13] J. Farrell and M. Rabin (1996), “Cheap Talk”, *The Journal of Economic Perspectives* **10**, 103-118.

[14] N. Kartik, M. Ottaviani, and F. Squintani (2007), “Credulity, Lies, and Costly Talk”, *Journal of Economic Theory* **134**, 93–11 6.

[15] S.A. Matthews, M. Okuno-Fujiwara, and A. Postlewaite (1991), “Refining Cheap-Talk Equilibria”, *Journal of Economic Theory* **55**, 247-273.

[16] Pei-yu Lo (2007), “Language and Coordination Games”, unpublished manuscript.

[17] R.A. Weber, C.F. Camerer and M Knez (2004), “Timing and Virtual Observability in Ultimatum Bargaining and “Weak Link” Coordination Games”, *Experimental Economics* **7**, 25–48.

[18] R. Zultan (2012), “Timing of messages and the Aumann conjecture: a multiple-selves approach”, *International Journal of Game Theory.*