1993

Cheap Talk and Signaling Games

Gyu Ho Wang

ISSN:0318-725X
ISBN:0-7714-151.

RESEARCH REPORT 9310

Cheap Talk and Signaling Games

by

Gyu Ho Wang

April 1993

Department of Economics

Social Science Centre

University of Western Ontario

London, Ontario, Canada

N6A 5C2

Cheap Talk and Signaling Games

Gyu Ho Wang

Department of Economics, The University of Western Ontario,

London, Ontario, Canada N6A 5C2.

April 1993

Abstract

We examine the role of cheap talk in the class of signaling games. For this, we define extended signaling games which extends the signaling games by allowing the player with private information to talk before he sends a costly signal. We propose a criterion called "Credibility Test" by which when cheap talk should be taken seriously. We show that "Credibility Test" is stronger than perfect sequential equilibrium by Grossman and Perry (1986). In class of signaling games of coordination, "Credibility Test" picks up the full coordination as the unique equilibrium outcome. Sometimes "Credibility Test" is too strong to exist.

1 Introduction

Despite the importance of natural languages in information transmission, the issue of how a meaningful communication can occur through natural languages has been a relatively unexplored area in game theory. In the specific context of a sender-receiver cheap talk game where only the sender has private information and only the receiver can take payoff-relevant actions, Crawford and Sobel (1982) showed how a meaningful communication could arise through cheap talk: Although the sender and the receiver do not have identical preferences over the final outcome, through cheap talk some degree of coordination is made. Later, for the class of sender-receiver cheap talk games, Farrell (1995,1988, 1990) provided the first formal refinement based upon the literal meaning of natural languages, which he called "Neologism-proof equilibria". His main point is that the literal meaning of natural languages should be respected more often than not. He proposed a criterion by which whether a neologism, any unsent message in the equilibrium under consideration is credible, and argued that once the message turned out to be credible, the literal meaning should override the equilibrium meaning which is determined within the equilibrium. Farrell's point was further investigated in the class of sender-receiver games by several authors. For example, Rabin (1990) suggested "Credible Message Rationalizability", Mathews, Okuno-Fujiwara, and Postlewaite (1991) suggested several kinds of "Announcement-proof equilibria".

The purpose of the paper is that Farrell's point that a meaningful communication can arise by natural languages need not be restricted to the class of sender-receiver games. We attempt to extend his point to the class of signaling games. A signaling game proceeds as follows; An agent called player 1 learns his type. Then, he takes some costly action in order to signal his private information to the other agent called player 2. Once having received the costly signal, player 2 chooses responses, and the game ends. This simple class of games has been extensively studied by several authors, for example Banks and Sobel (1987), Cho and Kreps (1987), and Grossman and Perry (1986). In order to examine the role of natural languages in signaling game, we extend the signaling game by adding one more stage, called *extended signaling game*. In extended signaling game, after having learned his type and before taking

the costly action, player 1 is allowed to talk anything he wants to player 2. The talk is cheap in the sense that it does not affect the payoffs. So in extended signaling game, player 1 can communicate in two ways, by costless signal (or cheap talk) and costly signal. The costless signal and the costly signal have their own advantage and disadvantage in conveying information. The advantage of costless signal over costly signal is that since it is assumed that both players share a rich, natural languages, there is no difficulty in understanding the literal meaning. The costless signal is more direct in conveying information. On the contrary, the costly signal is rather indirect. Usually, the costly signal is nothing but some physical action. What player 1 try to say by a costly signal should be determined within the equilibrium. It does not have any fixed meaning. It has equilibrium meaning only, and its equilibrium meaning is very equilibrium-specific. The disadvantage of costless signal is that since it is costless, player 1 may babble because player 2 believes player 1 babbles, therefore player 2 does not believe at all what player 1 says. Then, no meaningful communication arises by costless signal. For costly signal, since it is costly to send it, player 1 does not send arbitrary costly signal. So once having received a costly signal, player 2 tries to understand what player 1 is willing to say by that specific costly signal. Because of their relative advantage and disadvantage, we view the costless and costly signal are complementary with each other. In particular, with aid of different costless signals, player 1 can send different information with one costly signal. This may enable more coordination which is not possible without costless signal.

The paper is organized as follows. Section 2 and 3 provide the formulation of signal games and extended signaling games. Section 4 begins with some motivating examples, and gives the formal criterion called "Credibility Test". We show that "Credibility Test" is stronger than perfect sequential equilibrium by Grossman and Perry (1986). It has the strongest cutting power in the class of signaling games of coordinations. In that class of games, "Credibility Test" picks up the full coordination as the unique equilibrium outcome. "Credibility Test" determines the credibility of pair of costless and costly signal with respect to the equilibrium under consideration. We show by example that "Credibility Test" may tend to require player 2 to believe the literal meaning of costless signal rather than the equilibrium meaning of costly signal. It may lead to the problem of non-existence. Our view is that "Credibility

Test" is *some* way that rational player 2 may accept the literal meaning of costless signal combined with costly signal. Whenever it predicts a reasonable outcome, we accept its predictions. Conclusion follows.

2 Signaling Games

A signaling game consists of three stages; *learning stage*, *signaling stage* and *response stage*. At learning stage, player 1 learns his *type* t, drawn from a finite set T according to some probability distribution π over T. At signaling stage, player 1, having learned his type, sends a costly signal[1] s to player 2 chosen out of some finite set S. We allow the set of costly signals available to player 1 to depend on his type. We write $S(t)$ for the set of costly signals available to type t, and $T(s)$ for the set of types that have available the costly signal s. At response stage, player 2, having received a costly signal, chooses a response r from a finite set of responses R. We also allow the available responses to depend on the costly signal received, writing $R(s)$. The game ends with this response, and payoffs are made to the two players, depending upon the type of player 1, the costly signal player 1 sent, and the response player 2 took. The payoffs to player 1 and 2 are denoted by $u(t,s,r)$ and $v(t,s,r)$, respectively. We denote a signaling game by G:

$$G = \{T, \pi, S, R, u, v\}, \text{where } S = \cup_{t \in T} S(t),\ R = \cup_{s \in S} R(s).$$

The game structure, G is assumed to be common knowledge between player 1 and 2.

We specify strategies. We write behavior strategies for player 1 as $\sigma_1(s;t)$, where $\forall t \in T$, $\sigma_1(\cdot;t)$ is a probability distribution over $S(t)$. According to the behavior strategy $\sigma_1(\cdot;\cdot)$, player 1 with type t sends the costly signal s with probability $\sigma_1(s;t)$. We write behavior strategies for player 2 as $\sigma_2(r;s)$, where $\forall s \in S$, $\sigma_2(\cdot;s)$ is a probability distribution over $R(s)$. According to the behavior strategy $\sigma_2(\cdot;\cdot)$, having received s, player 2 chooses response r with probability $\sigma_2(r;s)$.

Rationality dictates that when player 2 has received a costly signal s, he should choose a response r which is a best response to *some* posterior belief μ over $T(s)$. We

[1]In the literatures on signaling games, the costly signal is usually referred to as *message*. We reserve the term message for costless signal, i.e., the message refers to *cheap talk*.

write $BR(\mu, s)$ for the set of best responses for player 2 to costly signal s if player 2 has posterior belief μ:

$$BR(\mu, s) = \text{argmax}_{r \in R(s)} \sum_{t \in T(s)} v(t, s, r)\mu(t).$$

For subset K of $T(s)$, let $BR(K, s)$ denote the set of best responses for player 2 to posterior beliefs whose support lies in K:

$$BR(K, s) = \cup_{\{\mu:\mu(K)=1\}} BR(\mu, s).$$

We denote by $MBR(\mu, s)$ and $MBR(K, s)$ the set of mixed best responses by player 2, respectively to belief μ and any belief whose support is in K. We denote a system of beliefs by $\mu(\cdot;\cdot)$ which describes the posterior beliefs by player 2. $\forall s \in S, \mu(\cdot; s)$ is a probability distribution over $T(s)$ which represents the posterior belief held by player 2, having received s.

A sequential equilibrium is a pair of a strategy profile $\sigma = (\sigma_1, \sigma_2)$ and a system of beliefs $\mu(\cdot;\cdot)$ which satisfies following two conditions:

1. Sequential Rationality:

For player 1, given σ_2, $\forall t \in T, \sigma_1(\cdot; t)$ maximizes the expected payoff. Namely, if $\sigma_1(s; t) > 0$, $s \in \text{argmax}_{s \in S(t)} \sum_r u(t, s, r)\sigma_2(r; s)$. For player 2, given $\mu(\cdot;\cdot)$, $\forall s \in S, \sigma_2(\cdot; s)$ maximizes the expected payoff when the posterior beliefs are given by $\mu(\cdot;\cdot)$; $\forall s \in S, \sigma_2(\cdot; s) \in MBR(\mu(\cdot; s), s)$.

2. Consistency:

When a costly signal s is sent with positive probability, $\mu(\cdot; s)$ should be computed using Bayes' rule; i.e., if $\sum_{t' \in T(s)} \sigma_1(s; t')\pi(t') > 0$,

$$\forall t \in T(s),\ \mu(t; s) = \frac{\sigma_1(s; t)\pi(t)}{\sum_{t' \in T(s)} \sigma_1(s; t')\pi(t')}.$$

Usually the consistency requirement is stronger than the Bayesian updating. In signaling game, however, due to its simple structure, the consistency requirement does not put any further restrictions on the out-of-equilibrium beliefs. In other words, player 2 can assign any posterior belief to the unsent costly signal. The plethora of out-of-equilibrium beliefs has been the main source of refining the equilibrium

in signaling games, for example, Cho and Kreps (1987), Banks and Sobel (1987), Grossman and Perry (1986).

Conditional on that type t is realized, a sequential equilibrium (σ, μ) induces interim payoffs for player 1 and 2 as follows:

$$u(t) = \sum_s \sum_r u(t,s,r)\sigma_1(s;t)\sigma_2(r;s) \text{ and } v(t) = \sum_s \sum_r v(t,s,r)\sigma_1(s;t)\sigma_2(r;s).$$

We call the array of payoffs $(u(t), v(t))_{t \in T}$ the *equilibrium interim payoffs.*

In signaling games, by sending a specific costly signal, to some extent player 1 tries to convey his private information to player 2, aiming at inducing *some* response by player 2. As a mere physical action, a costly signal does not have meaning as itself. What player 1 try to say to player 2 by a specific costly signal is determined within the equilibrium in effect. Within a given equilibrium, what player 1 means to player 2 by a specific costly signal is summarized in the posterior belief it induces. So we may call the posterior belief which a particular costly signal induces its *equilibrium meaning.*[2] Two things should be emphasized. First, a costly signal does not have any *fixed* meaning. It has equilibrium meaning only and the equilibrium meaning is very equilibrium-specific. Depending upon the equilibrium in effect, the same costly signal may have different equilibrium meanings. Second, within an equilibrium in effect, each costly signal, whether it is used or not in that equilibrium, has only *one* equilibrium meaning. In other words, within the fixed equilibrium, player 1 cannot communicate two different meanings with one costly signal.

3 Extended Signaling Games

An *extended signaling game* is obtained by adding to a signaling game one more stage called *talking stage* between learning stage and signaling stage. At talking stage, having learned his type, player 1 can talk anything he wants to player 2. Then, he sends a costly signal. The talk is cheap in the sense that regardless what player 1 says, it does not affects the payoffs directly. Although the talk is cheap, to the extent

[2]Of course, if a costly signal is sent with positive probability in a particular equilibrium, its equilibrium meaning in that equilibrium should be determined by Bayes law.

that the preferences are similar, player 1 may try to deliberately convey his private information with cheap talk and support its literal meaning by a costly signal, or may simply babble. At least, adding talking stage increases more opportunity for communication.

In order for some meaningful communication to arise at talking stage, it should be the case that there is no misunderstanding of the literal meaning of cheap talk. For this, we assume that player 1 and player 2 share a set of rich, natural languages whose literal meaning is clearly understood between them. Formally, we denote by M the set of natural languages shared between two players. We assume M is countable. Any element $m \in M$ is called a *costless signal*, *cheap talk* or *message*. So we use costless signal, cheap talk and message interchangeably.

In extended signaling games, the behavior strategies for player 1 has larger ranges and the behavior strategies for player 2 has larger domain. By slight abuse of the notation, we still denote a behavior strategy for player 1 by σ_1. In extended signaling game, $\sigma_1(\cdot, \cdot; t)$ is a probability distribution over $M \times S(t)$.[3] For behavior strategies for player 2, $\sigma_2(\cdot; m, s)$ is a probability distribution over $R(s)$ defined for all pair of (m, s). The system of beliefs also has a larger domain. $\mu(\cdot; m, s)$ should be defined for all pair of (m, s)

[3]According to the interpretation of the extended signaling game, the following formulation of the behavior strategy for player 1 may be more suitable: At talking stage, player 1 chooses a *talking strategy* $\rho(\cdot; t)$, a probability distribution over M. According to $\rho(\cdot; \cdot)$, player 1 with type t sends a costless signal m with probability $\rho(m; t)$. Then, at signaling stage, he chooses a *signaling strategy* $\phi(\cdot; m, t)$, a probability distribution over $S(t)$. A behavior strategy for player 1 is a pair of talking strategy and signaling strategy. This formulation may fit better the interpretation of the extended signaling game. However, note that a signaling strategy should be defined for all pair of message and type although a message is not used by the type in talking strategy. This complicates the notations only without any further generality. In order to avoid unnecessary complication, we define a behavior strategy such that depending on his type, player 1 chooses a joint distribution over $M \times S(t)$ with the following interpretation. Player 1 chooses m first according to the *marginal distribution* over M, and once m is chosen, send the costly signal according to the *conditional distribution* over $S(t)$ given m.

Since the definition of a sequential equilibrium in signaling games can be easily extended to the extended signaling games, we omit the definition. The equilibrium interim payoffs are defined similarly.

As is well-known, the addition of talking stage usually expands the set of equilibria. At least, a sequential equilibrium in the signaling game remains as a sequential equilibrium in the extended signaling game.[4] This expansion of set of equilibrium provides us with an opportunity to refine the equilibrium based upon the *literal meaning* of the costless signal. As is well-known, however, given a sequential equilibrium in extended signaling game, we can always construct another sequential equilibrium which uses all the messages in M and induces the same distribution over terminal nodes. So faced with cheap talk, any formal refinement loses its cutting power. Although it does not formally contradict with game theory, it seems too extreme that every equilibrium uses all messages in M. As pointed out by Farrell (1985), under some circumstances, although talk is cheap, its literal meaning should be respected. In other words, the literal meaning of costless signal should override the equilibrium meaning of costly signal. The purpose of the paper is to provide a criterion in signaling games by which when the literal meaning of cheap talk should be respected. For this, following Farrell (1985) we assume that in any equilibrium, there is a large set of unsent messages.

4 Examples and Credibility Test

In this section, we provide some motivating examples and formal criterion called "Credibility Test". We begin with the following example, $G(1)$.
In $G(1), \pi(t_1) = \pi(t_2) = 1/2$.

[4]To be more precise, given a sequential equilibrium in the signaling game, there is a sequential equilibrium in the extended signaling game which induces the same distribution over terminal nodes. To see this, let $\sigma = (\sigma_1, \sigma_2)$ and $\mu(\cdot;\cdot)$ be a sequential equilibrium in signaling game. Choose any probability distribution $p(\cdot)$ over M. Define $\sigma_1'(m, s; t) = p(m) \cdot \sigma_1(s; t)$, $\sigma_2'(r; m, s) = \sigma_2(r; s)$ and $\mu'(\cdot; m, s) = \mu(\cdot; s)$. It can be easily checked that $\sigma' = (\sigma_1', \sigma_2')$ and $\mu'(\cdot;\cdot,\cdot)$ is indeed a sequential equilibrium in the extended signaling game which induces the same distribution as $(\sigma, \mu(\cdot;\cdot))$. In this equilibrium, at talking stage, player 1 simply babbles.

Figure 1 Inserted here

In $G(1)$, there are three sequential equilibria.

Type 1: $\sigma_1(s_2;t_1) = \sigma_1(s_2;t_2) = 1, \sigma_2(r_1;s_1) \in [1/3, 2/3], \mu(t_1;s_1) = 1/2$ with equilibrium interim payoffs (2,2) for both t_1 and t_2.

Type 2: $\sigma_1(s_1;t_1) = \sigma_1(s_2;t_2) = 1, \sigma_2(r_1;s_1) = 1, \mu(t_1;s_1) = 1$ with equilibrium interim payoffs (3,3) for t_1 and (2,2) for t_2.

Type 3: $\sigma_1(s_2;t_1) = \sigma_1(s_1;t_2) = 1, \sigma_2(r_2;s_1) = 1, \mu(t_2;s_1) = 1$ with equilibrium interim payoffs (2,2) for t_1 and (3,3) for t_2.

In $G(1)$, regardless of his type, player 1 and player 2 have the identical preferences over the terminal nodes. So they want to coordinate on (3,3). In $G(1)$, however, (3,3) for both t_1 and t_2 cannot be equilibrium interim payoffs because player 2 cannot take different response in the information set following s_1. In extended signaling game, however, (3,3) for both t_1 and t_2 can be equilibrium interim payoffs by following pair of strategy profile and a system of beliefs; Pick up $m_1, m_2 \in M$ with $m_1 \neq m_2$. Set $\sigma_1(m_1, s_1; t_1) = \sigma_1(m_2, s_1; t_2) = 1, \sigma_2(r_1; m_1, s_1) = 1$, $\forall m \neq m_1, \sigma_2(r_2; m, s_1) = 1, \mu(t_1; m_1, s_1) = 1, \forall m \neq m_1, \mu(t_2; m, s_1) = 1$.

We view (3,3) for both types are the most sensible equilibrium interim payoffs in the extended signaling game of $G(1)$. All 1 - 3 Type equilibria fail the following test. We consider Type 2 and 3 first. Since they are symmetric, we only consider Type 2. In Type 2 equilibrium, regardless of messages, the equilibrium meaning of s_1 is t_1 with probability 1, therefore player 2 chooses r_1 with probability 1. This makes t_2 send s_2. Since t_1 enjoys the highest possible payoff, he has no incentive to deviate from the equilibrium. Suppose t_2 deviates from the equilibrium as follows. He first sends a message m whose literal meaning is "I am type t_2." Then, he chooses s_1. Once having received (m, s_1), player 2 should understand that only t_2 has incentive to do so. Then, he should respect the literal meaning of the message. In other words, the literal meaning of message m should override the equilibrium meaning of costly signal s_1. Then, having received (m, s_1), he chooses r_2, which overturns the equilibrium.

In Type 1 equilibrium, regardless of messages, the equilibrium meaning of s_1 is that player 1 is of either type with probability 1/2 so that player 2 chooses r_1 with probability at least 1/3 and at most 2/3. In this equilibrium, both types will deviate.

t_1 deviates, send a message m_1 whose literal meaning is "I am type t_1." and chooses s_1. t_2 also deviates, send a message m_2 whose literal meaning is "I am type t_2." and chooses s_1. Once having received (m_1, s_1) (resp., (m_2, s_1)), player 2 should understand that it is only t_1 (resp., t_2) who has incentive to send (m_1, s_1) (resp., (m_2, s_1)). So both m_1 and m_2 should override the equilibrium meaning of s_1, which means player 2 chooses r_1 (resp., r_2), having received $((m_1, s_1)$ (resp., $(m_2, s_1))$. Then, Type 1 equilibrium collapses.

Consider the next example, $G(2)$. In $G(3), \pi(t_1) = \pi(t_2) = \pi(t_3) = 1/3$.

Figure 2 inserted here

In $G(2)$, there is a unique sequential equilibrium; $\sigma_1(s_1; t_1) = \sigma_1(s_3; t_2) = \sigma_1(s_3; t_3) = 1$. $\sigma_2(r_1; s_1) = \sigma_2(r'_3; s_2) = 1$. $\mu(t_1; s_1) = 1, \mu(t_2; s_3) = \mu(t_3, s_3) = 1/2$.

This equilibrium, however, is again overturned by following consideration. In this equilibrium, regardless of messages, the equilibrium meaning of s_1 is t_1 with probability one and the equilibrium meaning of s_3 is that player 1 is either t_2 or t_3 with probability $1/2$. Since t_1 enjoys the highest possible payoff, he has no incentive to deviate. However, both t_2 and t_3 will deviate. t_2 sends a message m_1 and chooses s_1. t_3 sends a message m_2 and chooses s_2. In (m_1, s_1), t_2 says, "I am t_2 and choose s_1, anticipating that you will respond with r_2. If I were t_3, I would rather say to you that I am t_3 and choose s_3, anticipating that you will respond with r'_1." Similarly, in (m_2, s_3), t_3 says, "I am t_3 and choose s_3, anticipating that you will respond with r'_1. If I were t_2, I would rather say to you that I am t_2 and choose s_1, anticipating that you will respond with r_2." Player 2 either believes both messages or none. He cannot believe one message without believing the other. If player 2 believes (m_1, s_1), but does not believe (m_2, s_3) so that he will respond to (m_1, s_1) with r_2, but to (m_2, s_3) with r'_3. Then, t_3 mimics t_1 rather than sending (m_1, s_1) because by sending (m_2, s_3) he gets 4, but if player 2 will respond to (m_1, s_1) with r_2, he gets 5. However, player 2 knows this incentive by t_3. So he cannot believe (m_1, s_1), either. He will respond to (m_1, s_1) with r_1. So in order to overturn the above equilibrium, it will be necessary that player 2 should believe both messages.[5] Our view is that both messages are credible so that player 2 should believe both messages simultaneously. In the extended signaling game

[5]Similar phenomenon is pointed out by Mathews, Okuno-Fujiwara and Postlewaite (1991) in the

of $G(2)$, there arises new equilibrium interim payoffs; (10,10) for t_1, (9,9) for t_2 and (8,8) for t_3. In this equilibrium outcome, full coordination arises for all types of player 1. In $G(2)$, we view the full coordination as the most intuitive outcome.

Now we provide the formal definition of the criterion called "Credibility Test"

Fix a sequential equilibrium (σ, μ) in the extended signaling game with $(u^e(t), v^e(t))_{t\in T}$ as equilibrium interim payoffs. We say (σ, μ) fails the "Credibility Test " if following is true:

1. $\exists$ non-empty set K, K' (possibly empty) and $\sigma_1'(\cdot,\cdot;t)$ for all $t \in K \cup K'$ such that

(a). $\forall t \in K$, $\sum_{(m,s)} \sigma_t'(m,s;t) = 1$, $\forall t \in K'$, $\sum_{(m,s)} \sigma_t'(m,s;t) \leq 1$.

(b). If $\sigma_1'(m,s;t) > 0$ for some $t \in K \cup K'$, $\forall t \in T, \sigma_1(m,s;t) = 0$.

2. $\forall (m,s)$ with $\sigma_1'(m,s;t) > 0$ for some $t \in K \cup K'$, there exist $\mu'(\cdot;m,s)$, a probability distribution over $T(s)$ and $\phi(\cdot;m,s) \in MBR(\mu'(m,s),s)$ such that

(a). $K = \{t \in T |\ \forall (m,s)$ with $\sigma_1'(m,s;t) > 0, (m,s) \in \text{argmax} \sum_r u(t,s,r)\phi(r;m,s) > u^e(t)\}$,

(b). $K' = \{t \in T |\ \forall (m,s)$ with $\sigma_1'(m,s;t) > 0, (m,s) \in \text{argmax} \sum_r u(t,s,r)\phi(r;m,s) = u^e(t)\}$.

3. $\forall (m,s)$ with $\sum_{t\in K\cup K'} \sigma_1'(m,s;t)\pi(t) > 0$,

$$\mu'(t;m,s) = \frac{\sigma_1'(m,s;t)\pi(t)}{\sum_{t'\in(K\cup K')\cap T(s)} \sigma_1'(m,s;t')\pi(t')}.$$ [6]

Some remarks are in order. (a) in Condition 1 requires the existence of types player 1, K and K' who wants to deviate from the equilibrium under consideration. In any equilibrium, when (m,s) is used with positive probability, its equilibrium meaning given by Bayes law should be respected. So the types of player 1 who wants to deviate cannot test equilibrium meaning of (m,s) in use. Instead, they challenge

---

context of sender-receiver game.

[6] In the context of sender- receiver game, Mathews, Okuno-Fujiwara and Postlewaite (1991) consider the similar test. The difference is that while their test requires that the defectors should gain against all best responses, our test requires that they should gain against some best response. For this, see footnote 7. Also, in signaling game, due to its richer structure than sender-receiver game, the literal meaning of a costless signal can be taken more seriously with one costly signal than with other costly signals. This feature is missing in sender-receiver game.

the equilibrium meaning of unsent pair of (m, s) with σ_1'. Condition 2 says, given (m, s) sent by σ_1', player 2 will form a posterior belief given by μ' and will respond with ϕ. Player 2 is rational because $\phi \in MBR(\cdot, \cdot)$.[7] Given ϕ, (a) requires that K is exactly the set of types who has strict incentive to deviate. Furthermore, (a) requires that $\forall t \in K, \forall (m, s)$ with $\sigma_1'(m, s; t) > 0$, (m, s) gives player 1 with type t the same, highest payoff which is strictly greater than the equilibrium payoff. We need this in order to prevent the incentive for one type to mimic the other among the types in K. Similar consideration applies to K'. (b) requires that K' is exactly the set of types who are indifferent between deviating and following the equilibrium. Player 1 of type t in K' will deviate with probability $\sum_{(m,s)} \sigma_2'(m, s; t)$ and follow the equilibrium with remaining probability. He may randomize because he gets the same payoff as the equilibrium payoff by deviation. Condition (3) says the posterior belief μ' is not arbitrary. It should be computed by Bayes law using σ_1'.

We view that if conditions (1) - (3) are satisfied, the equilibrium meaning of (m, s) by μ in the equilibrium under consideration should be overridden by μ', which means the equilibrium in consideration should collapse.

Among many refinements in signaling games, our test is similar in its spirit to perfect sequential equilibrium by Grossman and Perry (1986). The main difference is that while the most refinements in signaling games including perfect sequential equilibrium consider an unsent costly signal in isolation, our test considers several out-of-equilibrium pair of message and costly signal simultaneously. So it is harder to pass our test. Below, we prove that "Credibility Test" is stronger than perfect sequential equilibrium. In extended signaling game, as seen in $G(1)$, it is possible that depending upon the type, each type of player 1 who wants to deviate sends different information with one costly signal because the costly signal is supplemented by several costless signals with different literal meanings. Also as seen in $G(2)$, in order for our test to have a cutting power, it is necessary that player 2 should believe

[7] Bayesian Rationality dictates that rational agent should choose an action which maximize the utility given his belief. When there are multiple best actions, it does not say anything about which one to choose. We may strengthen (2) by requiring for *all* best responses rather than for *some* best response. We view this is too strong. Furthermore, if we insist on it, our criterion loses its refining power dramatically.

all the pairs of message and costly signal used by the types of player 1 who deviate.

Now we compare "Credibility Test" with some other refinements in signaling game. As we said earlier, "Credibility Test" is stronger than perfect sequential equilibrium.

Proposition 1. Let (σ, μ) be a sequential equilibrium in a signaling game. If it fails to be a perfect sequential equilibrium, it fails "Credibility Test" in extended signaling game.

Proof: Let $u^e(t)$ be an equilibrium payoff associated with (σ, μ) for player 1 of type t. If (σ, μ) fails to be a perfect sequential equilibrium, there is an unsent costly signal s, non-empty subset K and K' (possibly empty) and ρ_1 such that $\forall t \in K, \rho_1(t) = 1$ and $\forall t \in K', \rho_1(t) \in [0,1]$. Furthermore, when $\nu(\cdot)$, a probability distribution over $T(s)$ is given by

$$\forall t \in K \cup K', \nu(t) = \frac{\rho_1(t)\pi(t)}{\sum_{t \in K \cup K'} \rho_1(t')\pi(t')} \text{ and } \forall t \in T - (K \cup K'), \nu(t) = 0,$$

there is $\phi' \in MBR(\nu, s)$ such that $K = \{t \in T | \sum_r u(t,s,r)\phi'(r) > u^e(t)\}$ and $K' = \{t \in T | \sum_r u(t,s,r)\phi'(r) = u^e\}$. In order to show (σ, μ) fails "Credibility Test", choose an unused message m and define $\forall t \in K, \sigma_1'(m,s;t) = 1, \forall t \in K', \sigma_1'(m,s;t) = \rho_1(t)$, and $\mu'(\cdot; m, s) = \nu(\cdot)$ with $\phi(\cdot; m, s) = \phi'(\cdot)$. Then, it can be easily checked that by construction σ_1' and μ' satisfy conditions (1) - (3). So (σ, μ) fails "Credibility Test". Q.E.D.

The proof shows that if a sequential equilibrium fails to be a perfect sequential equilibrium, "Credibility Test" eliminates the equilibrium simply using one pair of message and costly signal. However, "Credibility Test" allows several pairs of message and costly signal in order to eliminate the equilibrium in consideration. "Credibility Test" is strictly stronger than perfect sequential equilibrium. In $G(1)$, both Type 2 and 3 are perfect sequential equilibria. But both fails "Credibility Test".

The theorem in Grossman and Perry (1986) shows that any sequential equilibrium which does not pass *Intuitive Criteria* in Cho and Kreps (1987) is not a perfect sequential equilibrium. So following is the corollary of Proposition 1.

Corollary If a sequential equilibrium in a signaling game fails to pass Intuitive Criteria, it also fails "Credibility Test" in the extended signaling game.

"Credibility Test" has the strongest cutting power in the class of signaling games of coordination to be defined below.

Definition We say a signaling game is a game of coordination if $\forall t \in T$, $\text{argmax}_{(s,r)} u(t,s,r) \cap \text{argmax}_{(s,r)} v(t,s,r) \neq \emptyset$, where argmax is taken over (s,r) with $r \in BR(T(s),s)$. Let $u^*(t)$ and $v^*(t)$ be the associated payoffs for player 1 and 2, respectively, when player 1 is of type t.

$G(1)$ and $G(2)$ are examples of signaling game of coordination. Signaling game of coordination is a game such that every type of player 1 has incentive to identify his type and player 2 has incentive to believe that. Note that for every type of player 1, any equilibrium payoff for him cannot exceed $u^*(t)$. However, even in signaling games of coordination, as seen in $G(1)$ and $G(2)$, generally we cannot obtain $(u^*(t), v^*(t))_{t\in T}$ as equilibrium interim payoffs. However, in extended signaling game of coordination, we can always obtain $(u^*(t), v^*(t))_{t\in T}$ as equilibrium interim payoffs. Furthermore, $(u^*(t), v^*(t))_{t\in T}$ is the unique equilibrium interim payoffs associated with any sequential equilibrium which passes "Credibility Test".

Proposition 2. In the class of signaling games of coordination, there is a sequential equilibrium in extended signaling game whose equilibrium interim payoffs are $(u^*(t), v^*(t))_{t\in T}$. Furthermore $(u^*(t), v^*(t))_{t\in T}$ is the unique equilibrium interim payoffs for any sequential equilibrium in extended signaling game which passes "Credibility Test".

Proof: We prove the existence first. $\forall t \in T$, let $(s_t, r_t) \in \text{argmax}_{(s,r)} u(t,s,r) \cap \text{argmax}_{(s,r)} v(t,s,r)$. Let's choose distinct m_t from M for each t. Define σ_1 as follows: $\forall t, \sigma_1(m_t, s_t; t) = 1$. Define σ_2 as follows: For all pairs of $(m_t, s_t), \sigma_2(r_t; m_t, s_t) = 1$. For all other pairs of (m,s), choose a probability distribution $\nu(m,s)$ over $T(s)$ and let $\sigma_2(r; m, s) = 1$ where $r \in BR(\nu(m,s), s)$. For system of belief, for all pairs of $(m_t, s_t), \mu(t; m_t, s_t) = 1$. For all other pair of (m,s), set $\mu(m,s) = \nu(m,s)$ chosen above. In $\sigma = (\sigma_1, \sigma_2)$, given type t, both player 1 and 2 get the highest possible payoffs so that no one has incentive to deviate. The system of belief μ also follows Bayes law. So (σ, μ) is a sequential equilibrium in extended signaling game whose equilibrium interim payoffs are $(u^*(t), v^*(t))_{t\in T}$.

Now we prove uniqueness. Let (σ, μ) be a sequential equilibrium with $(u^e(t), v^e(t))_{t\in T}$

as associated equilibrium interim payoffs. Suppose $u^e(t) \neq v^*(t)$ for some t. Since any equilibrium payoff for type t cannot exceed $u^*(t)$, if we let $K = \{t \in T | u^e(t) < u^*(t)\}$, then, K is non-empty. For each t in K, choose distinct unsent message m_t. Define $\forall t \in K, \sigma_1'(m_t, s_t; t) = 1$. Note that any type $t \in T - K$ at least weakly prefers the equilibrium payoff. So any type $t \in T - K$ has no strict incentive to mimic some type in K. So $\forall t \in T - K, \forall (m, s)$, set $\sigma_1'(m, s; t) = 0$. For all pairs of (m_t, s_t) with $t \in K$, let $\mu'(t; m_t, s_t) = 1$. Then σ_1' and μ' satisfies conditions (1) - (3) in "Credibility Test". Condition 1 is clearly satisfied. Since $K = \{t \in T | u(t, s_t, r_t) = u^*(t) > u^e(t)\}$, (a) in condition 2 is satisfied. In constructing σ_1', by setting $\forall t \in T - K, \forall (m, s), \sigma_1'(m, s; t) = 0$, we choose an empty set as K'. Since all types in $T - K$ at least weakly prefers the equilibrium payoff, (b) in condition 2 is also satisfied. Note that $\forall t \in K, r_t \in BR(\mu'(m_t, s_t), s_t)$. So Condition 2 is satisfied. Since μ' can be computed by Bayes law using σ_1', Condition 3 is satisfied, too. The existence of (σ_1', μ') shows that (σ, μ) fails "Credibility Test". Q.E.D.

As we see, "Credibility Test" is a very strong criterion. It is stronger than perfect sequential equilibrium, and in the class of signaling games of coordination, it picks up the best payoffs for both player 1 and player 2 as the unique equilibrium payoff. Sometimes it is too strong to exist. Consider the following example, $G(3)$. In $G(3)$, $\pi(t_1) = \pi(t_2) = 1/2$.

Figure 3 Inserted here

$G(3)$ is simple because for t_1 and t_2, s_2 is strictly dominated by s_1. So at its unique equilibrium, both t_1 and t_2 send s_1 and player 2 will respond with r_3. There is no further equilibrium in extended signaling game. Although the equilibrium meaning of signal is determined within the equilibrium in consideration, in $G(3)$, under *any* circumstance, the equilibrium meaning of s_1 should be "obvious". It should be that player 1 is of either t_1 and t_2 with probability 1/2, which means player 2 should respond to s_1 with r_3. However, this unique equilibrium does not pass "Credibility Test". According to "Credibility Test", if t_1 deviates, sending a message whose literal meaning is "I am t_1." and choosing s_1, player 2 should believe this and revise the posterior belief which puts probability 1 on t_1, thereby responding with r_1 because it is t_1 only who can benefit from this deviation. By mimicking t_1, t_2 surely loses

compared with the equilibrium payoff. Then, the equilibrium fails "Credibility Test". This example shows that sometimes "Credibility Test" may require player 2 to believe too easily the literal meaning of costless message rather than the equilibrium meaning of costly signal. As long as we maintain equilibrium analysis, it seems that in $G(3)$ the equilibrium meaning of s_2 should precede the literal meaning of any costless signal. We do not interpret this as saying that "Credibility Test" is completely unreasonable. We do not argue that "Credibility Test" is the only way the rational player 2 should take the literal meaning of costless signal combined with costly signal and it should apply in every case. Our position is rather moderate. "Credibility Test" is *some* way that rational player 2 may accept the literal meaning of costless signal combined with costly signal. Whenever it predicts a reasonable outcome, we accept its predictions.

5 Conclusion

In this paper, we restrict our concern to the class of signaling game and address the issue of how rational player will take the literal meaning of natural languages. Clearly, in our daily life, natural languages play a very important role in information transmission, thereby affecting the decision making by economic agent. How rational players interpret the literal meaning of natural languages in more general context of games remains a big open question for game theory to solve. This question requires further investigation.

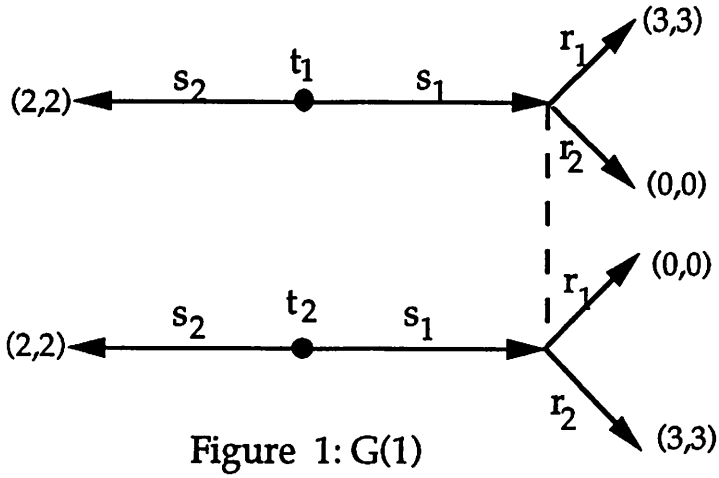

Figure 1: G(1)

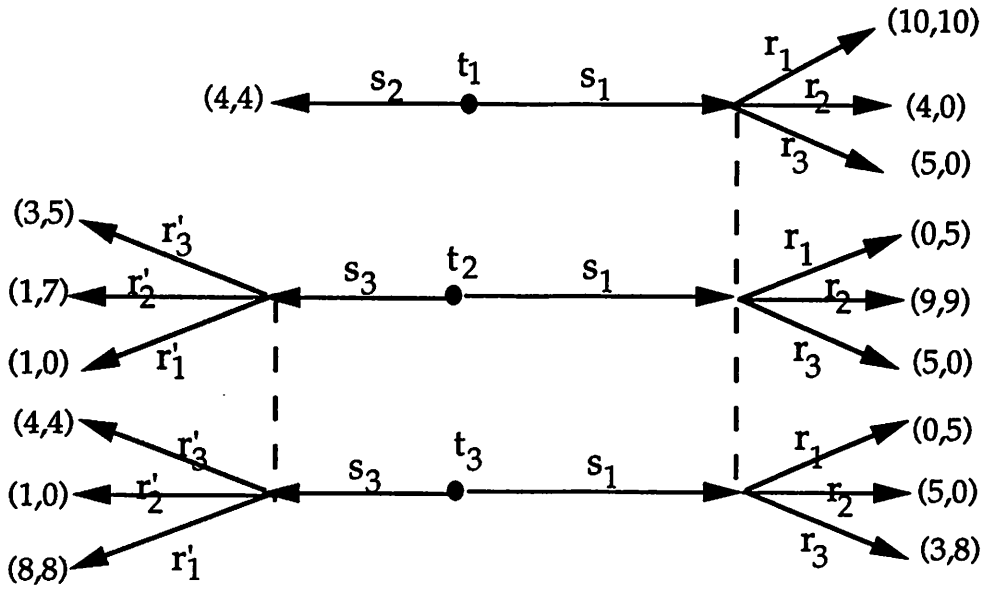

Figure 2: G(2)

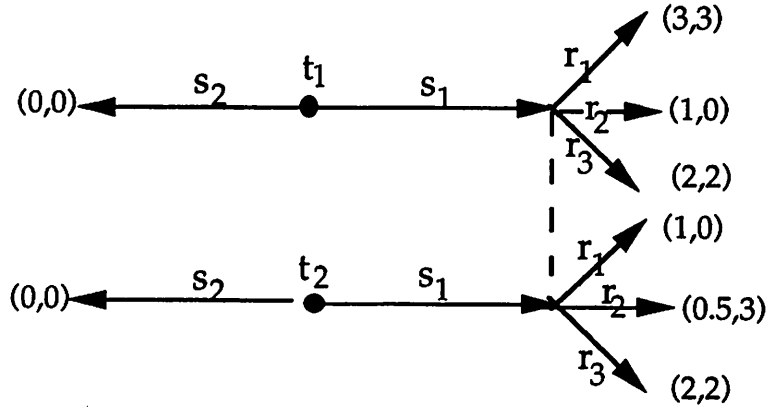

Figure 3: G(3)

References

[1] J. Banks and J. Sobel (1987), "Equilibrium selection in signaling games," *Econometrica* 55, 647-662.

[2] I.-K Cho and D. Kreps (1987), "Signaling games and Stable equilibria," *Quarterly Journal of Economics* 102, 179-221.

[3] V. Crawford and J. Sobel (1982), "Strategic information transmission," *Econometrica* 50, 1431-1451.

[4] J. Farrell (1985), "Credible Neologisms in Games of Communication," Working paper No. 386, MIT.

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

[6] J. Farrell (1990), Meaning and credibility in cheap-talk games, *in* "Mathematical Models in Economics" (M. Demster, Ed), Oxford University Press, London/New York.

[7] S. Grossman and M. Perry (1986), "Perfect sequential equilibrium," *Journal of Economic Theory* 39, 97-119.

[8] M. Rabin (1990). "Communication between rational agents," *Journal of Economic Theory* 51, 144-170.

[9] S. Mathews, M. Okuno-Fujiwara, and A. Postlewaite (1991), "Refining Cheap-Talk Equilibria," *Journal of Economic Theory* 55, 247-273.