Economics Letters 36 (1991) 375–378
North-Holland

Evolutionary stability in unanimity games with cheap talk

Karl Wärneryd *
Center for Study of Public Choice, George Mason University, Fairfax, VA 22030, USA

Received 25 January 1991
Accepted 13 May 1991

This paper studies a class of games of perfectly coinciding interests with a preplay communication stage added. If messages are costless (cheap talk), there are inefficient Nash equilibria even under communication. However, I show that only strategies that induce efficient outcomes are evolutionarily stable.

1. Introduction

That we cannot guarantee efficient outcomes in coordination games with perfectly coinciding interests is a major problem in game theory. It is especially troubling when we assume communication is possible. We have a strong intuition that communication solves coordination problems. Recently, this intuition has been used in an axiomatic fashion to refine Nash equilibrium [in, e.g. Farrell (1987, 1988) and Rabin (1990)]. There is, however, no accepted noncooperative foundation for this. The introduction of costless signals, or cheap talk, does not affect the set of Nash equilibria that can be induced in the underlying game.

Elsewhere [Wärneryd (1990)], I apply the concept of evolutionary stability to coordination games with imperfect information of the type studied in seminal contributions by Crawford and Sobel (1982) and Lewis (1986). In Kim and Sobel (1990) and Matsui (1991), more involved criteria based on evolutionary concerns are applied to games of the type to be studied here. In particular, Matsui shows that only efficient outcomes fulfill this criterion of cyclical stability in 2×2 games of coinciding interests. In this paper, I show that the simpler criterion of evolutionary stability [Maynard Smith (1982)] suffices.

The evolutionary stability point of view seems a natural one to take for purposes of explaining the effects of communication. Verbal utterances, or any other costless signals, can only be said to have specific meanings in relation to established conventions of interpretation. Such conventions necessarily arise out of historical processes of trial and error and imitation. While it does not get into the details of such processes, the criterion of evolutionary stability puts some reasonable restrictions on their stable states.

Section 2 introduces a class of 2×2 pure coordination games with cheap talk. I show that these games have inefficient equilibria, since cheap talk does not affect payoffs. In Section 3, I instead

* I thank the Swedish Council for Research in the Humanities and Social Sciences for financial support.

apply the notion of evolutionarily stable strategies. Only communication strategies that induce efficient outcomes fulfill the criterion of neutral stability. Finally, in Section 4, I note that this result does not generalize to games with more than two strategies.

2. The model

Consider the following 2×2 game with perfectly coinciding interests. The players share the action set $A = \{a_H, a_L\}$. The payoffs are the following.

	a_H	a_L
a_H	p_H, p_H	0, 0
a_L	0, 0	p_L, p_L

We will assume that $p_H > p_L > 0$. Pure coordination games of this type are sometimes called *unanimity games* [see, e.g., Harsanyi and Selten (1988)]. They can be seen as models of, for instance, simple bargaining situations.

The pure strategy equilibria of this game are (a_H, a_H) and (a_L, a_L). They are *strict* in the sense that they have no alternative best replies. Therefore the inefficient outcome cannot be ruled out according to any standard refinement criteria, including evolutionary stability.

Now extend the game by allowing each player to send a costless signal from the finite set Σ, with $|\Sigma| \geq |A|$, prior to actual play. The term *costless signal* is used to indicate that these actions:

(1) Do not affect the payoffs of the underlying game, and
(2) Do not have any particular meanings. [1]

In other words the communication stage consists of *cheap talk*.

Let $U = \Sigma \times \Sigma$ be a player's set of information sets reached after communication. Note that if a player is at the information set (σ, σ'), then his opponent is at the information set (σ', σ). A pure strategy for the game with communication is the choice of a signal σ to send and a function ϕ: $U \to A$. The function ϕ specifies an action for each of a player's possible information sets after the communication stage. Let $P(\sigma\phi, \sigma'\phi')$ be the payoff to a player using $\sigma\phi$ against an opponent playing $\sigma\phi$. Note that $P(\sigma\phi, \sigma'\phi) = P(\sigma'\phi', \sigma\phi)$.

Signals that do not affect payoffs and do not have conventional meanings in some language seem intuitively useless. It is easy to verify that the communication game has inefficient equilibria where both players choose strategies of the type $\sigma'\phi'$, where $\phi' = a_L$ for all $u \in U$. Although not strict, these equilibria involve undominated strategies and are therefore perfect in the sense of Selten (1975).

3. Evolutionary stability

A more natural approach to communication than the study of one-shot games is to ask which semantic conventions, i.e., systems of assigning meaning to signals, would be stable through repeated

[1] Signals will be said to be associated with meanings only in evolutionarily stable conventions. It would seem strange to assume signals can have meanings in history-less, one-shot interactions.

interactions in a large population of players. The concept of *evolutionary stability* [see, e.g., Maynard Smith (1982) or van Damme (1987)] offers a criterion based on such a viewpoint.

Definition 1 (evolutionary stability). Let Γ be a symmetric two-player game with strategy set $S = \{s_1, s_2, \ldots, s_n\}$ and payoff function P: $S^2 \to R$ that are the same for both players. A strategy $s' \in S$ is said to be an evolutionarily stable strategy (ESS) of Γ if, for all $s \in S$,

$$P(s^*, s^*) > P(s, s^*), \tag{1}$$

or, if $P(s^*, s^*) = P(s, s^*)$, then

$$P(s^*, s) > P(s, s). \tag{2}$$

The definition implies that an ESS is a symmetric Nash equilibrium strategy. In case the equilibrium is not strict, the strategy must fulfill an additional criterion. Note that both strategies of the game without cheap talk are ESS.

Although a criterion ostensibly defined for bimatrix games, we should really think of the ESS concept as a dynamic stability concept for populations. Consider a large population of identical players who meet randomly in pairs in each period to play the bimatrix game Γ. Let all players adhere to the same symmetric equilibrium strategy. Then the ESS criterion can be shown to be equivalent to requiring that no small group of deviators adhering to a different strategy can do better against the dominant strategy. In evolutionary terms, it guarantees that no such alternative strategy can grow over time, through, e.g., imitation, to take over domination of the population.

The ESS criterion is very restrictive. In fact, the cheap talk game does not have an ESS. However, the concept can be weakened in a reasonable manner to allow for *sets* of strategies that are collectively stable against deviation. It is done by replacing the strict inequality in (2) by a weak inequality. We call strategies fulfilling the weaker criterion *neutrally stable* strategies (NSS).

Proposition 1. A strategy $\sigma\phi$ is a NSS if and only if $(P(\sigma\phi, \sigma\phi) = p_H$.

To prove the 'if' part, let $\sigma\phi$ be such that $P(\sigma\phi, \sigma\phi) = p_H$. Let $\sigma'\phi'$ be an alternative best reply to $\sigma\phi$. We must have that $P(\sigma'\sigma', \sigma\phi) = P_H$. Since $P(\sigma\phi, \sigma'\phi') = P(\sigma'\phi', \sigma\phi) \geq P(\sigma'\phi', \sigma'\phi')$, $\sigma\phi$ is a NSS.

To prove the 'only if' part, let $\sigma\phi$ be such that $P(\sigma\phi, \sigma\phi) = p_L$. Let $\sigma' \neq \sigma$. Then if $\phi(\sigma, \sigma') = a_H$ we can find a ϕ' such that $\phi'(\sigma', \sigma) = a_H$. Then $\sigma'\phi'$ is a strict best reply to $\sigma\phi$, so $\sigma\phi$ is not a symmetric equilibrium strategy. On the other hand, if $\phi(\sigma, \sigma') = a_L$, we can find a ϕ' such that $\phi'(\sigma', \sigma) = a_L$ and $\phi'(\sigma', \sigma') = a_H$. Then $\sigma'\phi'$ is an alternative best reply to $\sigma\phi$ with $P(\sigma'\phi', \sigma'\phi') = p_H > p_L = P(\sigma\phi, \sigma'\phi')$, so $\sigma\phi$ is not a NSS.

Neutral stability allows for a set of strategies to be collectively stable against invasion by other strategies, but invadable by each other. This is not a problem, since the NSS sets contain strategies that differ only at information sets that are never reached in equilibrium.

The result may be interpreted as follows. Players come to interaction situations like this with a population history of previous play of similar games. If trial and error has gone on long enough to reach a stable state, there will exist a language that allows them to coordinate on the efficient solution.

4. Concluding remarks

The above result does not immediately generalize to unanimity games with more than two strategies. Consider the following example.

	a_H	a_M	a_L
a_H	p_H, p_H	0, 0	0, 0
a_M	0, 0	p_M, p_M	0, 0
a_L	0, 0	0, 0	p_L, p_L

Assume $p_H > p_M > p_L > 0$. The strategy $\sigma\phi$, where

$$\phi(u) = \begin{cases} a_M & \text{if } u = (\sigma, \sigma) \\ a_L & \text{otherwise,} \end{cases}$$

is a NSS. An alternative best reply to $\sigma\phi$ must send the signal σ and respond with a_M at (σ, σ). But then it will also play a_M against itself.

The positive result is really an instance of a more general theorem that excludes the *worst* outcome, rather than guaranteeing the best. The provision of more conclusive support for the intuition on communication and efficiency must be the task of future research in this area.

References

Crawford, Vincent P. and Joel Sobel, 1982, Strategic information transmission, Econometrica 50, no. 6, 1431–1451.
Farrell, Joseph, 1987, Cheap talk, coordination, and entry, Rand Journal of Economics 18, no. 1, 34–39.
Farrell, Joseph, 1988, Communication, coordination and Nash equilibrium, Economics Letters 27, 209–214.
Harsanyi, John C., Reinhard Selten, 1988, A general theory of equilibrium selection in games (MIT Press, Cambridge, MA).
Kim, Yong-Gwan and Joel Sobel, 1990, An evolutionary approach to pre-play communication, 1986, Manuscript.
Lewis, David K., 1986, Convention. A philosophical study (Basil Blackwell, Oxford).
Matsui, Akihiko, 1991, Cheap-talk and cooperation in a society, Journal of Economic Theory, fortcoming.
Smith, John Maynard, 1982, Evolution and the theory of games (Cambridge University Press, Cambridge).
Rabin, Matthew, 1990, Communication between rational agents, Journal of Economic Theory 51, 144–170.
Selten, Reinhard, 1975, Re-examination of the perfectness concept for equilibrium points in extensive games, International Journal of Game Theory 4, 25–55.
Van Damme, Eric, 1987, Stability and perfection of Nash equilibria (Springer, Berlin).
Wärneryd, Karl, 1990, Cheap talk, coordination, and evolutionary stability, Working paper (Center for Study of Public Choice, George Mason University, Fairfax, VA).