UNIVERSITY OF CALIFORNIA, BERKELEY

Department of Economics

Berkeley, California 94720

Working Paper 8609

MEANING AND CREDIBILITY IN CHEAP-TALK GAMES

Joseph Farrell

September 26, 1986

Key words: Communication, signaling, refinement of equilibrium, language, cheap talk, game theory.

There have been many previous and more technical versions of this paper; the most recent is listed in the references. For helpful comments and encouragement, I thank Vincent Crawford, Robert Gibbons, Jerry Green, Eric Maskin, Andrew McLennan, Roger Myerson, Suzanne Scotchmer, and especially Joel Sobel. I also thank two helpful referees, one of whom suggested the essay format.

JEL Classification: 026

ABSTRACT

In modeling verbal communication, it is natural to think of "messages" as not directly affecting payoffs: talk is cheap. Unfortunately, the standard restrictions on out-of-equilibrium beliefs scarcely if at all restrict beliefs in a model of cheap talk. This leaves us with an embarrassing plethora of equilibria.

If, instead of asking about equilibria in the game in isolation, we recognize the possibility that players share a rich natural language, then even messages not used in this equilibrium (neologisms) may have a focal meaning: their literal meaning. Although honesty may not always be the best policy, it is a focal policy, and we suppose that if there is no incentive to be dishonest, assuming that one's listeners assume one to be honest, then one will be honest: that is, speak the literal truth. This assumption links literal meaning to reality if it happens that there are no incentives to lie. In some cases, this restricts out-of-equilibrium beliefs, and hence restricts the set of equilibria. This refinement is the purpose of the paper.

There are three objections to this argument, which we discuss. First, every equilibrium outcome can be generated in an equilibrium in which all messages are used with positive probability; hence, arguments about out-of-equilibrium beliefs would seem to lack force. Second, how do neologisms have meaning? Third, why should a disequilibrium message be believed? We argue that, while these objections have some force, they do not completely meet the case.

We give examples showing what our proposed restriction on beliefs accomplishes, and note that there may be no equilibrium satisfying our restrictions. We discuss a dynamic evolutionary interpretation in which the absence of equilibrium means that the process never settles down.

1. INTRODUCTION

In a dynamic game of incomplete information, an informed player's actions may signal information. This idea is familiar to economists and game theorists when the cost of actions differs for different "types."[1] Crawford and Sobel (1982) introduced[2] information transmission without costs: an informed player may reveal information using costless "messages." This latter kind of communication, by words or cheap talk rather than by costly actions, is the topic of this paper.

Crawford and Sobel ask how informative an equilibrium language can be, given the degree of conflict and of common interest between the two players in a simple game. In general, they find many equilibria, and focus on the most informative because of its welfare properties. Game theorists are rightly suspicious of equilibrium-selection criteria that depend on ex-ante welfare arguments, so we try to use other techniques to pick out an equilibrium. But when talk is cheap, the usual selection techniques do not work: inferences from costless messages out of equilibrium are too arbitrary.

In this paper, we argue for a restriction on such inferences, and hence a refinement of the set of equilibria, in this cheap-talk case. We assume that the players already share a rich natural language, one that is much richer than is used in equilibrium in any particular game. Presumably this language comes from a history of diverse interactions and communication. When such a common language already exists, a message that is unexpected in a particular context may nevertheless be comprehensible. Its literal meaning is focal. Of course, when the players' interests do not completely coincide, that meaning may not be credible. But when (in a sense we make precise below) there is no incentive to lie, we assume that the unexpected message, or *neologism*, is believed. It turns out that the option of using such credible neologisms restricts the set of equilibrium outcomes, which was our goal.

The plan of the essay is as follows. In Section 2, we describe the importance of out-of-equilibrium beliefs in sequential equilibrium. We then define signaling games and cheap-talk games. In Section 3, we briefly describe some recent work on restricting out-of-equilibrium beliefs in signaling games, and explain why it does not apply to cheap-talk games. In Section 4, we argue that in equilibrium there will be unused possible messages. In Section 5, we discuss how a "neologism" or unexpected message may have meaning when it is costless. In Section 6, we ask about its credibility. In Section 7, we define the concept of neologism-proof equilibrium: one in which there are no credible unused messages that the sender would like to use. We give some examples. Section 8 briefly discusses an evolutionary interpretation of our argument. Section 9 concludes.

[1] For example, the choice of how much education to undergo may signal one's native ability [Spence (1974)]. Milgrom and Roberts (1982) model a monopolist's limit pricing by supposing that prices are taken to be signals of cost. Cramton (1984) and many other authors analyze how willingness to wait for a good price may signal reservation values in bargaining.

[2] See also Green and Stokey (1980), and Lewis (1969).

2. IMPORTANCE OF OUT-OF-EQUILIBRIUM BELIEFS

In a sequential[3] equilibrium, players' inferences from others' choices must satisfy Bayes' rule of rational inference. In an equilibrium in which every move is sometimes chosen, this requirement is enough to determine beliefs at every decision node. However, in a proposed equilibrium, there may be feasible moves that are never supposed to be chosen. If so, we must check that no player would wish to "defect" by choosing such a move.

Often, an important factor in a player's payoff from a move is the inferences that other players will draw from it; and in cheap-talk games this is his only concern. Therefore, it is essential to specify what the other players *would* infer from a move that supposedly will never be chosen.

Bayes' rule does not restrict players' inferences from such a zero-probability event. Accordingly, in sequential equilibrium, no restrictions[4] are placed on the theorist's freedom to specify such "out-of-equilibrium" beliefs. This freedom allows the theorist to make a deviation unattractive by specifying odious inferences, and so there are often many sequential equilibria. For example, in Spence's (1974) signaling model there are typically a continuum of them. Many theorists[5] find most of these equilibria implausible. In the next section, we discuss some restrictions on out-of-equilibrium beliefs that have been proposed to rule out these implausible equilibria. First, however, we define some terms.

A *signaling game* is as follows. An informed player, (the Sender, S) who knows the value of a random variable t in a set T, chooses a "message" m from a set M. Then an uninformed player (the Receiver, R) chooses an action a from a set A. Both players' payoffs depend on a, on S's true "type" t, and in general on m.

A *cheap-talk game* is a signaling game in which neither S's nor R's payoff depends on m: that is, payoffs are functions of a and t only.[6] In what follows, we focus on cheap-talk games. They are a natural model for ordinary verbal communication: talk is cheap. Because language is so important in human life and because the ability to talk often affects the outcomes of strategic interactions, this cheap-talk case is a very important one, even though it is of measure zero in the class of signaling games.

A message that is not used in an equilibrium is called a *neologism*.[7] In this essay, we discuss the meaning and credibility of neologisms, and the implications for equilibrium.

[3] We use this equilibrium concept [Kreps and Wilson (1982)] because it emphasizes beliefs and inferences. The limitations we shall discuss apply equally to perfect Bayesian equilibrium, for instance.

[4] In general, it is not quite true that no restrictions are placed on out-of-equilibrium beliefs in sequential equilibrium. However, the requirements of Kreps and Wilson's "consistency" have no force in the games we consider. (Roughly, consistency requires out-of-equilibrium beliefs to respect the constraint that players may not be able to distinguish certain decision nodes from others (imperfect information) and also requires that, "if possible," players should believe in only one defection rather than in two or more.)

[5] See, for instance, Riley (1979) or Kreps (1984).

[6] Crawford and Sobel (1982) call this a sender-receiver game.

[7] From the Greek for "new word." However, we think of neologisms as new (unexpected) messages or sentences composed in a common language, rather than as new words that have to be explained. But see Section 8 below.

3. STANDARD RESTRICTIONS ON SEQUENTIAL EQUILIBRIUM DO NOT APPLY TO CHEAP-TALK GAMES.

In the general signaling game, in which signals directly affect the sender's payoff, much recent work has investigated "reasonable" restrictions on out-of-equilibrium beliefs and the corresponding restrictions on equilibrium. Banks and Sobel (1985), Grossman and Perry (1985), Kreps (1984), and McLennan (1985) are important recent examples.[8] But none of these criteria rules out any sequential equilibrium outcomes[9] in cheap-talk games. All the restrictions are satisfied if we set each out-of-equilibrium belief equal to (some) belief that occurs in equilibrium. From an abstract point of view, this is quite reasonable: surely if one payoff-irrelevant action (in equilibrium) can induce a certain belief, then so can another, even though it was not meant to occur in equilibrium. But, as we argue below, this ignores the essential focal or coordinating ability of language, which can work even outside equilibrium.

However, the claim that out-of-equilibrium beliefs can be set equal to some equilibrium belief need not even be reached in "defending" a sequential equilibrium outcome, for any equilibrium outcome in a cheap-talk game can be supported by specifying that *all* messages are used in equilibrium. (To see this, begin with a proposed equilibrium. If some messages are unused, then pick any used message m, and reassign some of its probability weight to cover the formerly unused messages. Bayes' rule now tells us to interpret these messages in the same way as m.) Then the problem of out-of-equilibrium inferences does not even arise.

Thus an arbitrary sequential equilibrium outcome in a cheap-talk game has a three-layer defense against any attempt to rule it out by imposing natural restrictions on out-of-equilibrium beliefs.

[8] Kreps (1984) proposed to rule out a sequential equilibrium if there is a neologism m, a nonempty subset J of T, and a type t not in J, such that

(i) No type in J is willing to send m (he strictly prefers his equilibrium payoff) no matter what R would infer from m.

(ii) Type t strictly prefers sending m to using his equilibrium strategy, whatever R would infer from m provided that this inference excludes all types in J — that is, as long as by using m he could persuade R that he is not a type in J.

Intuitively, in such a case, the neologism m should convince R that t is not in J. This criterion (Kreps calls it "the intuitive criterion") rules out many implausible equilibria in generic signaling games. However, it has no force in cheap-talk games, since it is impossible to satisfy (i).

McLennan (1985) proposes ruling out (where possible) out-of-equilibrium beliefs that put positive weight on "useless" moves; i.e., those not used in any sequential equilibrium. But in a cheap-talk game, there are no useless moves (since language is arbitrary), and so this criterion too has no force.

Banks and Sobel (1985) start by observing that we can reasonably require R to infer from a neologism that S is of some type that *could* expect to benefit from deviating. That observation implies some restrictions on R's belief; and we can then ask what S-types might benefit given those restrictions; and so on. This leads to the concept of "divine" equilibrium. But again the restriction has no force in cheap-talk games.

Grossman and Perry (1985) propose that out-of-equilibrium as well as equilibrium beliefs should be "consistent" or "credible": that is, if possible, R should respond to a neologism m by finding a subset K of T such that precisely the members of K would benefit (nonstrictly, relative to equilibrium payoffs) by sending m if it convinced R that $t \epsilon K$. R should then believe that $t \epsilon K$. In a cheap-talk game, this condition does rule out some out-of-equilibrium beliefs; however, it allows all out-of-equilibrium beliefs to be equal to equilibrium beliefs, and therefore cannot rule out any equilibria.

First, every equilibrium outcome can be supported in such a way that all possible messages are used, and so there are no neologisms. Therefore, no argument concerning out-of-equilibrium beliefs can rule out any outcomes. Second, if "messages" are abstract payoff-irrelevant choices, then a neologism has no obvious meaning and an equilibrium meaning is as natural as any other; thus we can assign equilibrium meanings to any neologisms and get the same outcome. Third, even if there were neologisms that had some obvious meaning, it is not clear when such a neologism should be believed.

I shall argue that this defense is often misguided. In a cheap-talk game it is often plausible that there *are* unused messages with obvious meanings, and in some cases such messages are credible: that is, we can reasonably expect that their obvious meaning will be believed. Consequently, we can often rule out implausible sequential equilibria that are not ruled out by standard considerations.

We motivate our argument with an example of an implausible sequential equilibrium. Consider a game of pure coordination: two players must meet (choose the same place and the same time) in order to collect some payoff. It is intuitively clear that the outcome will be systematically different if they can talk first than if they cannot.[10] Yet there is always a sequential equilibrium (satisfying the criteria of Banks-Sobel, Grossman-Perry, Kreps, and McLennan) in which any messages are or would be ignored. We call this the uncommunicative equilibrium. It is not a plausible description of what happens if players can talk: a good theory will rule it out. We would like to find a reasonable restriction on beliefs in cheap-talk games that will rule out some implausible sequential equilibria, including the uncommunicative equilibrium in the case of pure coordination. That is the purpose of this paper.

[9] The *outcome* of an equilibrium is the function from types to (probability distributions on) the payoff-relevant actions chosen. Thus, it records the payoff-relevant aspects of the equilibrium while forgetting what costless messages are used.

[10] See the classic work of Schelling (1960).

4. ARE THERE UNUSED MESSAGES IN EQUILIBRIUM?

Given a fixed finite or countable[11] message space, it is formally possible to use all available messages with positive probability in an equilibrium, so that the question of out-of-equilibrium beliefs does not arise. However, there are several informal objections to this approach. While these objections are not conclusive — and sometimes there may indeed be no neologisms available — I believe that they carry enough weight to justify continuing to the next stage of the argument, in Section 5.

Intuitively, I want to argue that the set of possible messages is often open-ended, and so it is not possible to use all possible messages in a single equilibrium. Rather, there are always unused messages available. In an evolutionary interpretation of equilibrium (the game is played repeatedly with different participants), this is a very natural idea: there is no prior limit on the set of things that could be messages, and so there are always more signs that could convey information than do at any particular time. However, then neologisms do not immediately have focal meanings; a meaning must evolve. Accordingly, perhaps this interpretation is more suitable for analyzing the evolution of language by the introduction of neologisms than for refining our equilibrium prediction of how a game will be played. For more on this, especially the interpretation in terms of an evolutionary process, see Section 8 below.

In a one-shot framework it is hard to formalize the idea of open-endedness: if a message is possible, surely it can be used in equilibrium. Yet it is often implausible that, in fact, all messages are used in equilibrium to convey a few meanings. It would require the sender to randomize extensively, saying some very unnatural things, not for his own sake but for the sake of the equilibrium.

Perhaps the best way to capture the spirit of our argument is to stipulate that the sender, if possible, prefers to use messages that are short, simple, and expected to be interpreted straightforwardly. For example, if type t wants (and is expected) to reveal himself, and if both the English sentences, "I am t," and "I am either u or v," are interpreted in equilibrium as meaning t, then the sender will prefer the first. This suggests that it is hard to sustain mixed-strategy equilibria in which S randomizes over many messages with the same equilibrium meaning. If we rule out such randomization, and if T and A are both finite, then only finitely many messages will be used in equilibrium. Plenty will remain as neologisms.

In short, while there is no formal reason why all possible messages should not be used in equilibrium, it does not seem compelling as a way of sustaining an equilibrium outcome. We turn next to the question of meaning.

[11] If there are uncountably many messages in M, then it is not possible to attach positive probability to each one. This might seem to be a way out of the argument. But it is hard to insist that there are more than countably many messages; and even if we had a continuum, we might expect R to believe in zero-probability rather than in totally unexpected interpretations of what he hears. Thus, if a player is meant to name his "type," which is uniformly distributed on (0,1), and he says it is 0.123435875, R will presumably not treat that as a neologism, even though it was a probability-zero event.

5. WHY SHOULD A NEOLOGISM HAVE MEANING?

A message may have meaning in one of three ways. First, a meaning may be established by use [Wittgenstein (1958)]. This is the case for messages that are used in equilibrium: their meaning is established by Bayes' rule, which tells us the meaning-in-use of a message. Second, it may have a meaning that can be determined, or at least somewhat restricted, by introspection. This covers the restrictions on out-of-equilibrium beliefs discussed above, but they do not apply to cheap-talk games. Finally — and this is absent in previous analyses — a message may have a *focal meaning*.

Because this concept is unfamiliar to many game theorists who are used to thinking of the first two classes of meaning, an example may be useful. When the American revolutionaries wanted to signal how the British forces were coming, they agreed in advance that one light would mean "by land," two "by sea." If the British had come by air, or by tunnel, or if they had come both by land and by sea, three lights would not readily have conveyed the meaning, but the English (or American) language could have: the phrase "They're coming in balloons!" would have had a focal meaning (that they were coming in balloons).

This illustrates the difference between a prearranged set of meanings appropriate for the anticipated strategies in a given game, and a preexisting rich natural language. It is much like the difference between a *code* and a *cipher:* in a code, a list of possible meanings is fixed in advance and (cryptic) messages are chosen to convey those meanings. There are no meaningful neologisms. By contrast, a cipher is usually cryptically isomorphic to a natural language such as English. A much larger variety of meanings can be communicated — including the unanticipated, whether the surprise is exogenous or is a deviation from a proposed equilibrium.

We also see here the difference between our emphasis and that of Crawford and Sobel. If we ask what language structures can be equilibria in a particular game, considered in isolation, then it is reasonable to suppose (at least in a one-shot framework) that meaning is conferred only by established use: neologisms have no meaning. But when there is a rich common language, even a neologism (a message quite unexpected *in this context)* may be comprehensible: its *literal meaning* is common knowledge. Nothing *requires* players to take the literal meaning seriously, but it is focal and so a player might be wise to do so — if he believes that the other player is doing so. In some games, such as zero-sum games, the existence of a focal meaning is not useful: if the receiver knew what the sender wanted him to believe, he would not believe it. But in other cases, where their interests sufficiently coincide, he would.

What meaningful neologisms are available? The spirit of this essay is that every possible meaning can be conveyed (though it need not be believed). However, anticipating the discussion of credibility of neologisms in the next section, we need only a limited set of neologisms. We suppose that a credible neologism can only be sent *instead of* the equilibrium message (though alternative assumptions could be considered), and we also suppose, since we are checking an equilibrium for incentives to deviate, that only those types who strictly benefit from sending a neologism will do so. Since of course all such types will do so, it follows that, for any neologism, no type will "mix" sending and not sending. In that case, if a neologism is credible, it must claim that the sender is in some subset X of types, and R's posterior belief will be the restriction to X of his prior.

So we can assume without loss that meaningful neologisms take the form n(X): "My type is some $t \epsilon X$," for some nonempty subset X of T. We assume that *for every nonempty subset X of T, and for every sequential equilibrium of the game, there exists* a *message n(X) that is unused in the equilibrium* and *whose literal meaning is that* $t \epsilon X$.

For simplicity, we also assume (as is generically true if both T and A are finite) that R's best response a(X) to that belief is unique, for all nonempty subsets X of T. Thus, credibility is the only barrier to communication, out of equilibrium as well as in. In the next section, I propose a criterion for credibility of a neologism.

6. WHEN IS A MEANINGFUL NEOLOGISM CREDIBLE?

What would R infer from a meaningful neologism $n(X)$? He could infer that $t \in X$, but in general that would be very credulous. He should presumably consider what types of S might expect to do better than their (putative) equilibrium payoffs. Perhaps he should infer that S is one of the types that would prefer that R believe that $t \in X$ [and so play $a(X)$], rather than get their equilibrium payoff; we might denote this conclusion as $t \in P(X)$. Or he could go a step further and infer that this is what S would like him to believe, so that he should instead infer that $t \in P[P(X)]$. The multiple-bluff story can be extended as far as we like. Or should R put some probability on each of these possible inferences? In general, it seems unclear.

In one case, however, I suggest it is clear what R should believe. This is the case $P(X) = X$; we call such a subset X *self-signaling*, because S's desire to have R believe that $t \in X$ signals precisely that $t \in X$. S would like R to believe his message $n(X)$ if and only if it is true. We therefore suggest that if X is self-signaling[12] then the neologism $n(X)$ is credible: R should believe it.

When S chooses his message in an equilibrium, he has a choice of inducing in R any of the following beliefs: (i) all beliefs that R holds in equilibrium, and (ii) any other beliefs that R would hold out of equilibrium. Equilibrium is restricted if category (ii) is nonempty. Our assumption is that category (ii) contains, at least, any self-signaling sets X (more precisely, the restrictions to any such sets X of R's prior). This contrasts with the standard theory of cheap-talk games, in which it is admissible to make (ii) empty by definition (all out-of-equilibrium beliefs can be set equal to some equilibrium beliefs). This is the fundamental assumption of this essay. Before pursuing its implications, we pause to discuss two natural counter-arguments:

(i) In some cases, if R would interpret the absence of the neologism $n(X)$ to mean that $t \in T \setminus X$, then $n(X)$ is no longer self-signaling: it may be that the set of types who would prefer the action $a(X)$ to the action $a(T \setminus X)$ is not equal to X. Then, the argument goes, since everyone knows that $n(X)$ is available, it is not clear that it should be taken to mean that $t \in X$.

I believe that this argument is inconsistent with the notion of equilibrium in game theory. A proposed equilibrium that offers scope for profitable defection is not rescued by the fact that the profitable defection would be unprofitable if anticipated. For instance, in the game if (U,L) were proposed as an equilibrium, we should object that Row would defect to D. This defection would be unprofitable if Column anticipated it (he would then play R), but that does not make (U,L) an equilibrium.

	Column	
Row	**L**	**R**
U	(1,1)	(1,1)
D	(2,0)	(0,1)

(Row's payoff first)

[12] This depends not only on X but also on the proposed equilibrium payoffs.

Similarly, in testing whether a sequential equilibrium is neologism-proof, we should consider the consequences of an *unexpected* deviation (neologism). Both the payoff from the deviation and the payoff from the proposed equilibrium strategy should be calculated on the assumption that the deviation is unexpected; and the failure-to-occur of an unexpected event should not lead R to revise his beliefs. We do not propose an equilibrium in which n(X) is used, any more than we propose an equilibrium in the game just given in which D is used (there is none); but the possibility of profitable *unanticipated* deviation rules out an equilibrium.

(ii) There may be two self-signaling neologisms [say n(X) and n(Y)] available in a proposed equilibrium. Then we can ask whether the use of n(X) and not n(Y) should be interpreted in the same way as if n(Y) were not available. As in (i), one can argue that it should. In checking a proposed equilibrium, we assume that R does not expect deviations, and so he will not infer anything from a failure to use n(Y).

But one could argue that once he observes a deviation, R should reevaluate everything, including the other available deviations. His beliefs about the conduct of the game have been shattered; it might be wise for him to think the whole thing out afresh. In particular, although he is inclined to find n(X) convincing, he might ask what other equally convincing neologisms might have been used instead.

This argument leads to a somewhat different theory of credibility. For instance, one might deem a neologism n(X) credible if it is self-signaling and if no other self-signaling neologism n(Y) would give any S-type $t \in X$[13] a higher payoff than n(X). A referee suggests calling such a neologism *truly credible*.

Other definitions of credibility are possible. For example, one might insist that S name a whole new sequential equilibrium in which the types in X are treated as a group and in which precisely the set X of types is better off. This takes to the extreme the argument that all players should "anticipate" a neologism. Myerson's (1983) notion of *core mechanism* requires not only that S name a whole new equilibrium that is better for types in X, but also that the improvement work whether R indeed infers that $t \in X$, or makes no inference, or anything "in between."

[13] If X and Y are disjoint and both n(X) and n(Y) are self-signaling, then n(X) and n(Y) are also truly credible: for if n(X) is better than n(Y) for some $t \in Y$, then $t \in P(X)$ as well as in P(Y), which is impossible since $X = P(X)$ and $Y = P(Y)$ are disjoint. Therefore, only overlapping self-signaling neologisms will give trouble of this kind.

7. NEOLOGISM-PROOF EQUILIBRIUM

If there is a credible[14] neologism in an equilibrium, and if S has a clear incentive to use it, then the equilibrium is not self-enforcing. We say that such an equilibrium is not *neologism-proof.*

If self-signaling neologisms are credible, then they will be used, for by definition S strictly wishes to use such a neologism whenever it is true. The other criteria for credibility discussed above also have this property that any credible neologism will be used. Thus the very existence of a credible neologism makes an equilibrium not neologism-proof.

What equilibria are neologism-proof? We address this question through examples. In some cases, unreasonable-seeming sequential equilibria are ruled out, while the reasonable ones are neologism-proof. Perhaps less appealingly, we also find that no neologism-proof equilibrium need exist.

For our examples, we use the following assumptions and notation. There are two types of sender: A and B. The receiver has three different actions: a(A) is best for him when he is sufficiently confident that S is of type A, a(B) when S is of type B, and a(T) is best when the receiver has (close enough to) the prior probabilities in mind. We give in table form the payoffs to the two S-types when R takes each of his three actions.

Example 1: In this example, the players' interests coincide. In this case, the uncommunicative equilibrium is not neologism-proof.

Action	Payoff to A	Payoff to B
a(A)	2	0
a(B)	0	2
a(T)	1	1

There are two sequential equilibrium outcomes. In one, S reveals his type, and R takes the appropriate action a(A) or a(B). In the other, all messages are uninformative,[15] and R always chooses a(T).

As discussed in Section 3 above, standard considerations do not rule out this latter (uncommunicative) equilibrium. Neologism-proofness does so. The neologism n(A) is self-signaling, as is the neologism n(B).[16]

In this example, both players are better off in the unique neologism-proof equilibrium than in the uncommunicative equilibrium. In general, however, S need not be better off ex-ante. To see this, change the payoff to B from a(B) to −10, and suppose that A and B are equally likely ex-ante. Of course, R is always better off with more information, but it is possible to construct

[14] As emphasized above, a neologism's credibility depends on the putative equilibrium payoffs to the different S-types. Thus, this is a condition on the proposed equilibrium.

[15] Strictly, it is not necessary that R's posterior after any message should always be his prior, but only that his posterior never place enough weight on either type to justify his choosing the actions a(A) or a(B).

[16] Since these neologisms do not overlap, and since it is still desirable to identify oneself even if the absence of a neologism will be taken as significant, these neologisms are "truly credible."

an example with three types in which R does not have unambiguously more information in the neologism-proof equilibrium than in another sequential equilibrium, and ex-ante is worse off.

Example 2: Here, the separating equilibrium is not neologism-proof.

Action	Payoff to A	Payoff to B
a(A)	1	0
a(B)	0	1
a(T)	2	2

Here again, there are two sequential equilibria. In this case, however, it is the separating equilibrium that fails to be neologism-proof: the neologism n(T) is self-signaling (relative to that equilibrium). Intuitively, the content of n(T) is "I won't tell you my type. Since it is preferable for me whatever my type that you should not be confident about my type, you should not infer anything about my type from my refusal to disclose."

To support the separating equilibrium, such a message would have to be interpreted as (sufficiently strong)[17] evidence in favor of one type or the other. This seems to require some power of commitment on R's part. For example, if the separating equilibrium is very good for R, he might try to commit himself to "take" anything except the claim that $t \in A$ as indicating that $t \in B$. But unless there is an explicit understanding, we expect that n(T) will be effective in making R take the action a(T).

Example 3: In this example, there is no neologism-proof equilibrium. While type A wishes to distinguish himself from type B, type B prefers to be mistaken for a type A rather than identified as a type B. Thus, whenever the two types are treated alike, there is a self-signaling neologism; but there is no equilibrium in which they are treated differently.

Action	Payoff to A	Payoff to B
a(A)	2	1
a(B)	−1	0
a(T)	0	2

There is just one sequential (or Nash) equilibrium outcome: all equilibrium messages are uninformative,[18] and R always chooses action a(T). However, the neologism n(A) is then self-signaling. Thus no equilibrium is neologism-proof if self-signaling neologisms are credible.[19]

[17] That is, strong enough to make R willing to choose one of his "confident" actions a(A) or a(B).

[18] More precisely, none is sufficiently informative that R becomes confident enough to prefer a(A) or a(B) to a(T).

[19] n(A) is also "truly credible": that is, there is no competing self-signaling neologism. Therefore, that more-restrictive theory of credibility does not solve the existence problem. In this example, the equilibrium is preserved if R would require S to name a new equilibrium before being convinced by a neologism. However, it is possible to construct another example (three types are necessary) in which existence fails even then.

To sustain the sequential equilibrium in this example, it is necessary that R's posterior after any message (equilibrium or not) should induce either the action a(T) or the action a(B). If we believe, however, that a type-B would not say, "Really, I'm type-A; notice that I wouldn't want you to believe that if I weren't," unless all other messages were taken to mean type-B, then the only solution is to specify that R expects from both types an eloquent claim that t = A; if he hears anything less, he infers that t = B. This specification has the following unappealing property: any deviation can strictly benefit only A, but is assumed to mean B. Except for strict versus nonstrict inequalities, these out-of-equilibrium beliefs violate Kreps' (1984) "intuitive criterion."

If we change A's payoff under a(B) to +1, then the equilibrium requires that all messages (in and out of equilibrium) induce a(T). As argued in Section 4 above, we would not realistically expect to find messages such as "Honestly, I'm an A; please believe me," used by B if messages like "I won't tell you my type, in accord with equilibrium," also induce a(T). So the equilibrium then is even less plausible.

8. EVOLUTIONARY INTERPRETATION OF NEOLOGISMS

We have discussed neologisms in a one-shot game when there is already a rich common language. An alternative interpretation of equilibrium is as an "evolutionarily stable outcome," in which no "mutation" will grow in the population [Maynard Smith (1982)]. In this interpretation, it is natural to suppose that, while there are plenty of previously unused signs that could serve as messages, they do not convey meaning when first used. Thus the meaning of a neologism must evolve. How can this happen?

To illustrate, consider Example 3. Suppose that initially there is no communication. All S's are blondes and have blue eyes. Now suppose that, by chance, a few predominantly type-A S's develop red hair. There is then selective pressure on R's to respond to redheads with the action a(A), while (at first) continuing to use a(T) for blondes. Once a significant proportion of R's behave like that, there is strict selective pressure on type-A's to develop red hair, while the reverse is true for type-B's. Thus red hair will come to be a better and better signal of type-A, both in the sense that most type-A's have it and in the sense that most redheads are type-A.

At some point, however, enough type-A's will be redheads that it will pay for R to treat blondes as type-B's. As the proportion of R's who do so increases beyond 1/2, it becomes attractive for type-B's to become redheads: the alternative is no longer mostly a(T), which they prefer, but mostly a(B). As more type-B's become redheads, that signal degrades: eventually almost all S's of both types are redheads and the fact no longer conveys meaning. As R's adjust to that, we return to where we started: everyone is treated with a(T). Eventually, perhaps, it will happen that some R's (mostly A's) will develop brown eyes, and the story begins again.

We see that a self-signaling neologism is evolutionarily successful at first for precisely the types it claims, as long as their alternative continues to be the previous equilibrium treatment. Once that is no longer so, the signal may "degrade": in this example, it degrades by imitation by type-B's.

Thus in dynamic "equilibrium," there is sometimes revelation of type (constantly being eroded by imitation), and sometimes pooling (liable at any time to erosion from the appearance of neologisms). The average outcome will depend on the relative speeds of innovation and of imitation.

This is somewhat analogous to the situation in military science, in which it has been claimed that every offensive weapon can be defensively countered, but at any given time there may be offensive weapons whose defenses have not yet been developed. Likewise, a human who has a cold acquires an immunity to that cold virus, but if the community is large enough that the virus can rapidly develop new strains, then we are infected again. Although the body is good at developing immunity to any given cold virus, it cannot anticipate all the possible mutations.

This seems a reasonable description of what might happen in a game such as Example 3 in which there is no (static) equilibrium. The point of this heuristic story is twofold: not only to give the "evolutionary" interpretation, but also to suggest that the lack of static equilibrium means that things will not settle down, not that no coherent prediction can be made.

9. CONCLUSION

In games in which an informed agent may reveal information using costless messages (cheap talk), standard refinements of the sequential-equilibrium concept do not apply. However, not all sequential-equilibrium outcomes are reasonable. To eliminate unreasonable equilibria, we consider out-of-equilibrium focal meanings: the literal meanings of unexpected messages in a natural language. Plausibly, some such messages are credible, and if such messages would be believed, this can eliminate some equilibria.

In some cases, indeed, no equilibria remain. We can conclude that we have no satisfactory positive theory in a one-shot game. Alternatively, we can think of an evolutionary interpretation, in which case the lack of equilibrium means simply that things will not settle down.

This paper argues for taking games in context, especially when analyzing the effects of communication. Language that could not survive in equilibrium in a particular game can nevertheless affect the outcome of the game.

10. REFERENCES

BANKS, J., and J. SOBEL, "Equilibrium Selection in Signaling Games," mimeo, CalTech and UCSD (1985).

CRAMTON, P., "The Role of Time and Information in Bargaining," *Review of Economic Studies 167*, 579-594 (1984).

CRAWFORD, V., and J. SOBEL, "Strategic Information Transmission," *Econometrica 50* (1982).

FARRELL, J., "Credible Neologisms in Games of Communication," mimeo, GTE Labs and MIT (1985).

GREEN, J., and N. STOKEY, "A Two-Person Game of Information Transmission," mimeo, Harvard (1980).

GROSSMAN, S., and M. PERRY, "Sequential Bargaining Under Asymmetric Information," *Journal of Economic Theory,* forthcoming (June 1986).

______ and ______, "Perfect Sequential Equilibrium," mimeo, Princeton (1985), *Journal of Economic Theory,* forthcoming.

KOHLBERG, E., and J.F. MERTENS, "On the Strategic Stability of Equilibria," mimeo, CORE (1982), *Econometrica,* forthcoming.

KREPS, D., "Signaling Games and Stable Equilibria," mimeo, Stanford (1984).

KREPS, D., and R. WILSON, "Sequential Equilibrium," *Econometrica 50*, 863-894 (1982).

LEWIS, D., *Convention,* Cambridge, Mass.: Harvard University Press (1969).

McLENNAN, A., "Justifiable Beliefs in Sequential Equilibrium," *Econometrica 53*, 889-904 (1985).

MAYNARD SMITH, J., *Evolution and the Theory of Games,* London: Cambridge University Press (1982).

MILGROM, P., and J. ROBERTS, "Limit Pricing and Entry Under Incomplete Information" *Econometrica 50*, 443-459 (1982).

MYERSON, R., "Mechanism Design by an Informed Principal," *Econometrica 51*, 1767-1798 (1983).

RILEY, J., "Informational Equilibrium," *Econometrica 47*, 331-359 (1979).

RUBINSTEIN, A., "A Bargaining Model with Incomplete Information About Time Preferences," *Econometrica 53*, 1151-1172 (1985).

SCHELLING, T., *The Strategy of Conflict,* Cambridge, Mass.: Harvard University Press (1960).

SELTEN, R., "A Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory 4*, 25-55 (1975).

SPENCE, M., *Market Signaling,* Cambridge, Mass.: Harvard University Press (1974).

WITTGENSTEIN, L., *Philosophical Investigations,* translated by G. Anscombe. Third edition, Oxford: Blackwell (1958).

Sep-86

RECENT ISSUES OF THE WORKING PAPER SERIES OF THE DEPARTMENT OF ECONOMICS UNIVERSITY OF CALIFORNIA, BERKELEY

Copies may be obtained from the Institute of Business and Economic Research. See the inside cover for further details.

8601 Jeffrey A. Frankel
THE DESIRABILITY OF A CURRENCY DEPRECIATION, GIVEN A CONTRACTIONARY MONETARY POLICY AND CONCAVE SUPPLY RELATIONSHIPS
Feb-86.

8602 Drew Fudenberg and David M. Kreps
REPUTATION AND MULTIPLE OPPONENTS I: IDENTICAL ENTRANTS
May-86.

8603 Jeffrey A. Frankel and Kenneth A. Froot
EXPLAINING THE DEMAND FOR DOLLARS: INTERNATIONAL RATES OF RETURN AND THE EXPECTATIONS OF CHARTISTS AND FUNDAMENTALISTS
Jun-86.

8604 Jeffrey A. Frankel and Kenneth A. Froot
USING SURVEY DATA TO EXPLAIN STANDARD PROPOSITIONS REGARDING EXCHANGE RATE EXPECTATIONS
Aug-86.

8605 Jerry A. Hausman and Paul A. Ruud
SPECIFYING AND TESTING ECONOMETRIC MODELS FOR RANK-ORDERED DATA WITH AN APPLICATION TO THE DEMAND FOR MOBILE AND PORTABLE TELEPHONES
Aug-86.

8606 Roger Craine
RISKY BUSINESS: THE ALLOCATION OF CAPITAL
Aug-86.

8607 Leo K. Simon and Maxwell Stinchcombe
EXTENSIVE FORM GAMES IN CONTINUOUS TIME PART I: PURE STRATEGIES
Aug-86.

Sep-86

RECENT ISSUES OF THE WORKING PAPER SERIES
OF THE DEPARTMENT OF ECONOMICS
UNIVERSITY OF CALIFORNIA, BERKELEY

Copies may be obtained from the Institute of Business and Economic Research. See the inside cover for further details.

8608 Robert M. Anderson
THE SECOND WELFARE THEOREM WITH NONCONVEX PREFERENCES
Sep-86.

8609 Joseph Farrell
MEANING AND CREDIBILITY IN CHEAP-TALK GAMES
Sep-86.

8610 Joseph Farrell and Garth Solaner
COMPETITION, COMPATIBILITY AND STANDARDS:
THE ECONOMICS OF HORSES, PENGUINS AND LEMMINGS
Sep-86.