5
Meaning and credibility in cheap-talk games

Joseph Farrell
University of California, Berkeley

Chapter 5. *Mathematical Models in Economics*. MOL Bacharach, MAH Dempster, and JL Enos, eds. University of Oxford (1990).

5

Meaning and credibility in cheap-talk games

Joseph Farrell
University of California, Berkeley

Abstract

I define *neologism-proofness*, a refinement of perfect Bayesian equilibrium in cheap-talk games. It applies when players have a pre-existing common language, so that an unexpected message's literal meaning is clear, and only credibility restricts communication. I show that certain implausible equilibria are not neologism-proof; in some games, no equilibrium is.

1. Introduction

In a dynamic game of incomplete information, it is a familiar idea that an informed player's actions may signal information if the direct costs or benefits of actions differ for different 'types'.[1] But there is also information transmission without costs:[2] an informed player may reveal information using costless messages or *cheap talk*.

Two problems limit the effectiveness of cheap talk among selfish rational agents. The first, on which most game-theoretic attention has focused, is that of *credibility*: communication cannot work well when there are incentives to lie. In most games of 'mixed motive', in which participants' interests are partly common and partly conflicting, this constraint limits the equilibrium effectiveness of cheap talk. Crawford and Sobel (1982) show just how informative an equilibrium language

Thanks are due to Joel Sobel and Bob Gibbons, who encouraged me when editors, referees, and colleagues did not. Thanks are also due to the National Science Foundation for financial support (IRI-87-12238).

[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, analyse how willingness to wait for a good price may signal reservation values in bargaining.

[2] See Crawford and Sobel (1982), Green and Stokey (1980), and Lewis (1969).

can be, given the degree of conflict and of common interest between the two players in a simple game. Loosely, the informativeness of the *most*-informative equilibrium is limited by the degree to which the players' interests coincide. There are always other equilibria in which language is less informative. For instance, there is always a 'babbling' equilibrium, in which all messages are taken to be meaningless. Crawford and Sobel focus on the most-informative equilibrium in order to answer the question, 'what constraints on communication are imposed by the possibility of lies?'

But there is a second, more fundamental, problem in using cheap talk to communicate: its meaning cannot be learned from introspection. This immediately implies one, unimportant, kind of multiplicity of equilibrium in a cheap-talk game: any permutation of messages across meanings gives another equilibrium.

More technically but much more importantly, however, it implies that the multiplicity of perfect Bayesian equilibria cannot be reduced by standard refinement arguments. Intuitively, we can place no *a priori* restrictions on the interpretation of messages that were not expected in equilibrium ('neologisms'), and this fact disarms the refinements that argue from how certain unexpected messages 'should' be interpreted. Viewing an equilibrium (including the language actually used) in isolation, neologisms can logically be taken as having any arbitrary meaning, or none, or as pointless verbal variations on equilibrium messages.

But we need not view the language associated with an equilibrium in isolation. If the players share a rich common language such as English (developed in a much larger environment than this single game), then although only certain messages will be used in an equilibrium, players can say anything they like, and can expect that, although neologisms may not be *believed*, they will at least be *understood*. And certain neologisms, once understood, are intrinsically credible. Requiring that such 'credible neologisms' would be believed if uttered, imposes a condition on equilibrium—that no such credible neologism is available and attractive relative to the equilibrium—which we call 'neologism-proofness.' In what follows, we argue in more detail the case for this refinement, and investigate its implications.

The plan of the essay is as follows. In §2, we describe the importance of out-of-equilibrium beliefs in perfect Bayesian equilibrium. We then define signalling games and cheap-talk games. In §3, we briefly describe some recent work on restricting out-of-equilibrium beliefs in signalling games, and explain why it does not apply to cheap-talk games. In §4, we argue against the possible assumption

that in equilibrium all possible messages will be used (leaving no language in which to utter neologisms). In §5, we discuss how a neologism can have meaning. In §6, we ask about its credibility. In §7, we define a neologism-proof equilibrium: one in which there are no attractive credible neologisms. We illustrate and explore the concept using examples in §8. Section 9 discusses an evolutionary interpretation of our argument. Section 10 gives some general results. Section 11 considers applications; §12 concludes.

2. The importance of out-of-equilibrium beliefs

In a perfect Bayesian equilibrium, players' inferences from others' choices must satisfy Bayes's rule. In an equilibrium in which every move is sometimes chosen, this requirement determines beliefs after each possible history of the game. But in some perfect Bayesian equilibria, there may be feasible moves that are never chosen; and an equilibrium condition is that such moves are unattractive. Often, an important factor in a player's pay-off from a move is the inferences that others 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 in equilibrium is not chosen. As Cho and Kreps (1987, p.180) write,

> ... what constitutes an equilibrium is powerfully affected by the 'interpretations' that would be given by B to messages that A *might* have sent, but in equilibrium *does not* send.

Bayes's rule does not restrict these interpretations. Accordingly, in describing a perfect Bayesian equilibrium, the theorist is completely free to specify the 'out-of-equilibrium' beliefs—in particular, to specify odious inferences—and consequently there are typically many perfect Bayesian equilibria (for example, Spence's (1974) signalling model typically has a continuum of equilibria). Many theorists 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 *signalling game* is a simple two-player, two-stage game of incomplete information, as follows. An informed player (the Sender, S) who has privately observed a random variable $t \in T$, chooses a 'message' $m \in M$. Then an uninformed player (the Receiver, R) chooses an action $a \in A$. Both players' pay-offs depend on a, on t, and in general on m.

A *cheap-talk game* is a signalling game in which neither S's nor R's pay-off depends on m: that is, pay-offs are functions of a and t only. 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 intuitively 'of measure zero' in the class of signalling games.

A message that is not used in an equilibrium is a *neologism.*[3] Obviously, the property of being a neologism is only defined relative to a given equilibrium.

3. Standard refinements do not help for cheap-talk games

In the general signalling game, in which signals directly affect the sender's pay-off, much recent work has investigated 'reasonable' restrictions on the interpretations of out-of-equilibrium messages, and the corresponding restrictions on equilibrium.[4] But none of these refinements limits the set of equilibrium outcomes (that is, equilibrium functions from types t to probability distributions on pay-off-relevant actions a) in cheap-talk games. We next describe why this is so: there are two related reasons.

The first reason is that every such outcome is also the outcome of an(other) 'noisy' equilibrium: that is, one in which *all messages are used with positive probability.* To see this, consider any perfect Bayesian equilibrium: this consists of a (probabilistic) function from type t to messages m, and a (probabilistic) function from messages m to actions a. In this equilibrium, there may be some unused messages in M: let M_0 be the set of such messages. Now construct another equilibrium, with the same outcome, as follows. Take an arbitrary message $m^* \in M \setminus M_0$, and let S behave as follows: 'every time I would have sent the message m^*, I will now randomize over $M_0 \cup \{m^*\}$ in such a way that every $m' \in M_0 \cup \{m^*\}$ gets positive weight.' Since now the likelihood function on T conditional on any message $m' \in M_0 \cup \{m^*\}$ is identical with what it was conditional on m^* in the old equilibrium, it is an equilibrium for R to respond to all such messages in the same way as he did to m^* in the old equilibrium. And if R does so, then it is a best response for S to do as we have suggested. We have thus constructed a noisy equilibrium with the same outcome function as the original equilibrium. In this new equilibrium, there are no neologisms

[3] From the Greek for 'new word'. (According to the *Oxford English Dictionary*, the term dates from 1803.)

[4] See especially Banks and Sobel 1987; Cho and Kreps 1987; Grossman and Perry 1986*b*; Kohlberg and Mertens 1986; McLennan 1985. For an illuminating survey, see Cho and Kreps 1987.

whose reasonable interpretation might rule out the outcome. Thus the original outcome cannot be eliminated by refinement techniques, if we accept the noisy equilibrium.

A second (related) reason why standard refinements have no bite in cheap-talk games is that even if the equilibrium is not noisy—there are unused messages in equilibrium—none of the standard refinement arguments bars us from choosing a message m^* used in the equilibrium and interpreting every neologism to mean the same as m^*. Because standard refinements are based on introspection and on common knowledge of rationality, they cannot rule out such an interpretation.

From an abstract point of view, this is quite reasonable: surely if one pay-off-irrelevant choice m can induce a certain belief about t, then so can another, m', even though m' was not meant to occur in equilibrium. But, as we argue below, this ignores the focal nature of neologisms in an existing common language, which can work even outside equilibrium.

4. Are there unused messages?

We showed above that given a fixed (finite or countable) message space M, any equilibrium outcome can be represented as the outcome of a 'noisy' equilibrium, in which all messages $m \in M$ are used with positive probability, so that there are no neologisms available and the question of out-of-equilibrium beliefs does not arise. In order to formulate a refinement of perfect Bayesian equilibrium, therefore, I must argue that noisy equilibria are not entirely plausible.

Suppose for instance that M is the English language, and consider the noisy babbling equilibrium in the pure coordination game. This requires that, with positive probability, S says 'I will be at home' when in fact he knows that he will be at Z's restaurant—indeed, that S's saying he will be at home tells R nothing about where S actually plans to be. Since the two players have completely common interests, I suggest that this is very unlikely. Even if S were extremely pessimistic about the chances of effectively communicating with R (suppose, for instance, that it was notorious that 'R never listens to what you say') he would be unlikely to behave like that; rather, he would remain silent, or would say, 'I have given up trying to communicate with you: you never listen', or something of the kind. Thus—I claim—the noisy babbling equilibrium is implausible. It requires S to randomize extensively, saying some very unnatural things, not for his own sake but for the sake of the equilibrium.

Perhaps we can capture the spirit of our argument more generally if we postulate that S prefers *where possible* to use messages that are short, simple, and straightforward. 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 'I am t', then S will prefer the former. 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, and plenty will remain available as neologisms if M is infinite. Perhaps this idea could be further formalized by viewing cheap-talk games as limits of games in which S has a slight preference for telling the truth.

Another objection to noisy equilibria is that the set of possible messages is often 'open-ended', so that it is *not* possible to use all messages in an equilibrium. In an evolutionary interpretation of equilibrium (the game is played repeatedly with different participants), 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.[5]

In short, while there is formally nothing wrong with noisy equilibria, they do not seem compelling as a way of sustaining an equilibrium outcome when (as is usually the case) M is large. We consider next the *meaning* of neologisms, assuming that neologisms exist.

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 (1957) urged 'Don't ask the meaning; ask the use.' The meaning of messages that are used in equilibrium, in particular, is established by Bayes's rule, which tells us their meaning-in-use. Secondly, it may have a meaning that can be determined, or at least somewhat restricted, by introspection. This yields the restrictions on out-of-equilibrium beliefs discussed above; but they do not apply to cheap-talk games. Finally—and this is the key element absent in previous analyses—a message may have a *focal meaning*, if it is phrased in a pre-existing language.

---

[5] However, then neologisms do not immediately have focal meanings: a meaning must evolve. Accordingly, perhaps this interpretation is more suitable for analysing 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 §8 below.

Because the relevance of this concept is unfamiliar, an example may be useful. When the American revolutionaries wanted to be able to signal how the British 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). For, although the rebels did not expect to have to send or interpret such a message, it was common knowledge that they knew what the word 'balloons' meant. The pre-existing language was rich enough not only to provide for the expected messages, but to convey unexpected ones too.

The difference between a pre-arranged set of meanings appropriate for the anticipated messages in a given equilibrium, and a pre-existing rich natural language, is 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 (like the aerial redcoats) or is a deviation from a proposed equilibrium.

We see here the essential difference between our emphasis and that of Crawford and Sobel. If we ask what language structures can be equilibria in a given 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 may be comprehensible: its *literal meaning* is common knowledge.

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, for our purpose, we need only a limited set of neologisms. We assume that *for every non-empty* $X \subseteq T$, *and for every perfect Bayesian equilibrium of the game, there exists a message* $n(X)$ *that is unused in the equilibrium and whose literal meaning is that* $t \in X$. Thus, messages are comprehensible and 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?

We have argued that cheap-talk games whose players share a rich, pre-existing common language should be analysed assuming that there are meaningful neologisms available: for every message X that S might want to convey, there is a neologism $n(X)$ whose literal meaning is that S's type t lies in X. But this is a big step from assuming that S can actually persuade R to *believe* what he is saying. Our methodology has been to separate the problems of meaning (comprehensibility) and of credibility; we turn now to credibility.

Nothing *requires* players to take a neologism's 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 games of conflict, such as zero-sum games, the existence of a focal meaning is irrelevant: if the receiver knew what the sender wanted him to believe, he would not believe it. But where players' interests sufficiently coincide, he would.

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 pay-offs. Perhaps he should infer that S is one of the types that would prefer that R believe that $t \in X$ (and so play his best response[6] $a(X)$), rather than get their equilibrium pay-off: 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 is unclear what he should do, and what we should model him as doing.

But I suggest it *is* clear what R should believe if $P(X) = X$.[7] We call such a subset X ***self-signalling***, because S's wish 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 assume that if X is self-signalling[8] then the neologism $n(X)$ is credible: R should believe it. It is important to note that this is not a claim about equilibrium: it is consistent with equilibrium for R to interpret neologisms in various ways, as discussed above. Rather, this 'should' is

---

[6] 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 non-empty $X \subseteq T$.

[7] Notice that this also implies $P(P(X)) = X$, and so on.

[8] This depends not only on X but also on the proposed equilibrium pay-offs.

a claim, based on informal introspection and reasonableness,[9] about what we believe is likely to happen. Correspondingly, as we argue next, we are unconvinced by equilibria in which it is posited that R responds to self-signalling neologisms $n(X)$ with beliefs other than $t \in X$.

7. Neologism-proof equilibrium

When S chooses his message in an equilibrium, he can induce in R any of the following beliefs: (i) all beliefs that R holds in equilibrium, and (ii) any other beliefs that (according to the full specification of the equilibrium) R would hold out of equilibrium. In the standard theory of cheap-talk games, it is admissible to make category (ii) empty (set all out-of-equilibrium beliefs equal to some equilibrium beliefs). We assume, by contrast, that category (ii) contains, at least, any self-signalling sets X (more precisely, the restrictions to any such sets X of R's prior). This restricts the set of equilibria.

If there is a credible neologism available 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*; the attractive credible neologism breaks the equilibrium.

If self-signalling neologisms are credible, then any available self-signalling neologism breaks the equilibrium: for by definition S strictly wishes to use such a neologism whenever it is true. (And since this fact is common knowledge, the equilibrium does not hold even if S's type t happens not to be in the set X.)[10] Thus the very existence of a credible self-signalling neologism makes an equilibrium not neologism-proof.

Before pursuing its implications, we pause to discuss two natural counter-arguments: that is, two possible arguments as to why R might 'reasonably' disbelieve a self-signalling neologism $n(X)$ and thus preserve the equilibrium.

(i) In some cases, if S expects that R would interpret the absence of the neologism $n(X)$ to mean that $t \notin X$, then $n(X)$ is no longer self-signalling: there is no reason why the set of types who would prefer the action $n(X)$ to the action $n(T \setminus X)$ should be 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 $t \in X$.

---

[9] I also ran a small number of informal classroom experiments to test this idea, and the results were consistent with my claim, although not conclusive.

[10] The other criteria for credibility discussed above also have this property that any credible neologism will be used.

But 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

		Column's Move	
		L	R
Row's Move	U	$(1,1)$	$(1,1)$
	D	$(2,0)$	$(0,1)$

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.

Similarly, in testing whether a perfect Bayesian equilibrium is neologism-proof, we should consider the consequences of an *unexpected* deviation (neologism).[11] Both the pay-off from the deviation and the pay-off 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-signalling 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 re-evaluate 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 would lead to a somewhat different theory of credibility. For instance, one might deem a neologism $n(X)$ credible if it

[11] Cho and Kreps (1987. p.203) argue similarly in their beer-quiche example.

is self-signalling and if no other self-signalling neologism $n(Y)$ would give any S-type $t \in X$ [12] a higher pay-off than $n(X)$. One might call such a neologism *truly credible.*[13]

give any S-type $t \in X$ [12] a higher pay-off than $n(X)$. One might call such a neologism *truly credible.*[13]

8. Examples

We set out originally with two goals: first, to consider dispassionately what will happen in cheap-talk games when a rich common language is available in which to formulate neologisms; and secondly, to find a principled reason to dispose of some equilibria (such as the babbling equilibrium in the coordination game) that we found distasteful. What equilibria are neologism-proof? We address this question through examples. In some cases, unreasonable-seeming perfect Bayesian 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

---

[12] If X and Y are disjoint and both $n(X)$ and $n(Y)$ are self-signalling, 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 $t \in P(X) \cap P(Y)$, which is impossible since $X = P(X)$ and $Y = P(Y)$ are disjoint. Therefore only overlapping self-signalling neologisms will give trouble of this kind.

[13] Other definitions of credibility are possible. For example, one might insist that S name a whole new perfect Bayesian 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'.

[12] If X and Y are disjoint and both $n(X)$ and $n(Y)$ are self-signalling, 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 $t \in P(X) \cap P(Y)$, which is impossible since $X = P(X)$ and $Y = P(Y)$ are disjoint. Therefore only overlapping self-signalling neologisms will give trouble of this kind.

[13] Other definitions of credibility are possible. For example, one might insist that S name a whole new perfect Bayesian 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'.

the receiver has (close enough to) the prior probabilities in mind.[14] We give in table form the pay-offs to the two S-types when R takes each of his three actions.

Example 1: *Pure coordination.* In this example, the players' interests coincide. The uncommunicative (babbling) equilibrium is not neologism-proof. This perhaps vindicates our natural distaste for it.

R's Action	Pay-off to A	Pay-off to B
$a(A)$	3	0
$a(B)$	0	3
$a(T)$	2	2

[14] It is easy but tedious to write down primitive pay-offs for R and S from the different types and actions that lead to the data we give.

There are two perfect Bayesian 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 above, standard considerations do not rule out this implausible latter equilibrium; but neologism-proofness does so: the neologism $n(A)$ is self-signalling, 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. But this is not the point: in general, S need not be better off *ex ante* in the neologism-proof equilibrium. To see this, change the pay-off 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 the neologism-proof equilibrium need not be more informative:

Example 2: *I won't tell you.*

R's Action	Pay-off to A	Pay-off to B
$a(A)$	1	0
$a(B)$	0	1
$a(T)$	2	2

Here again, there are two perfect Bayesian equilibria. In this case, however, it is the separating equilibrium that fails to be neologism-proof: the neologism $n(T)$ is self-signalling (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, that neologism would have to be interpreted as (sufficiently strong)[17] evidence in favour of one type or the other. This seems to require some commitment on R's part. For example, if the separating equilibrium is good for R, he might try to commit himself to 'take' anything except the claim that

---

[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.'

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

$t = A$ as indicating that $t = B$. But unless there is an explicit meta-discussion in advance, we expect that $n(T)$ will make R choose $a(T)$. And if S believes that, the equilibrium is undone. ■

Example 3: *No neologism-proof equilibrium.* 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-signalling neologism; but there is no (perfect Bayesian) equilibrium in which they are treated differently.

R's Action	Pay-off to A	Pay-off to B
$a(A)$	2	1
$a(B)$	–1	0
$a(T)$	0	2

There is just one perfect Bayesian (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-signalling. Thus no equilibrium is neologism-proof.[19]

To sustain the perfect Bayesian 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 non-strict inequalities, these out-of-equilibrium beliefs violate Cho and Kreps' (1987) 'intuitive criterion'.

Indeed, if we change A's pay-off under $a(B)$ to $+1$, then the equilibrium requires that all messages (in and out of equilibrium) induce $a(T)$. As argued in §4 above, we would not realistically expect to find messages such as 'Honestly, I'm an A; please believe me',

---

[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-signalling 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 in which the subset X of T is separated from $T \setminus X$, before being convinced by a neologism $n(X)$. However, one can construct another example (three types are necessary) in which existence fails even then.

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. ■

9. Evolutionary interpretation

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: there is no pre-existing common language rich enough to communicate neologisms. 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 S's, by chance predominantly type A, 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, by chance some R's (mostly A's) will develop brown eyes, and the story begins again.

Thus the self-signalling neologism is evolutionarily successful 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.[20]

In military science, it has been claimed that every offensive weapon can be defensively countered, but at any given time there may be offensive weapons whose defences 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 story is to suggest that the lack of equilibrium means that things will not settle down, not that no prediction can be made.

10. Results

In this section, we give some results on existence and characterization of neologism-proof equilibrium.

Our first result concerns games in which S's preferences over R's beliefs are independent of S's true type: formally, for all $t_1, t_2 \in T$ and for all $a_1, a_2 \in A$, t_1 strictly prefers a_1 to a_2 if and only if t_2 does so. Intuitively, cheap talk will be ineffective in such games. Typically, the unique perfect Bayesian equilibrium is uncommunicative: that is, R's action is independent of S's type t in equilibrium.

Fact 1. *If S's preferences over R's beliefs are independent of t, then the uncommunicative perfect Bayesian equilibrium is neologism-proof.*

Proof. For any non-empty subset $X \subseteq T$, $P(X)$ must be either empty or the whole set T. Consequently, T is the only possible self-signalling subset. But in the uncommunicative equilibrium, $n(T)$ is not a neologism: it is used in the equilibrium. ∎

Fact 2. *In a game of pure conflict (including zero-sum games), the uncommunicative equilibrium is neologism-proof.*

Proof. Suppose that $X \subseteq T$ were self-signalling in the uncommunicative equilibrium. Then (from the definition) S strictly prefers $a(X)$ to $a(T)$ if $t \in X$. Since the game is one of pure conflict, this means that R strictly prefers $a(T)$ to $a(X)$ if $t \in X$. But this contradicts the definition of $a(X)$. ∎

[20] Dawkins and Krebs (1979) discuss the effect of selective pressure on relative speeds of adaptation in evolution.

Fact 3. *If R and S have identical interests (the case of pure coordination), then full communication is a neologism-proof equilibrium. If the type space T contains three or fewer elements, then full communication is the unique neologism-proof equilibrium. When T contains four elements, this need not be so.*

Proof. First, observe that no action of R's can be strictly better for type t than R's best response $a(\{t\})$, since S's interests are perfectly aligned with R's. Consequently, nothing is self-signalling in the full-communication equilibrium. This proves the first claim.

For the rest of the proof, see Farrell (1985). ∎

Corollary to Fact 3. *Neologism-proof equilibrium need not be unique if it exists.*

11. Applications

Gertner *et al.* (1988) analyse a firm's communication (not through cheap talk) of its profitability by a firm to a rival and to the capital market. They show that there may be many perfect Bayesian equilibria, but that there is a unique neologism-proof equilibrium. This need not be a separating equilibrium, and (as they point out) it is one of the features of neologism-proofness that it can pick out the pooling equilibrium (as in Example 2 above), while standard refinement concepts typically insist on 'at least' a certain amount of separation.

Farrell and Gibbons (1988*a*,*b*) show how cheap talk can matter in bargaining; in (1988*b*) they show that it *must* matter: that is, for some parameter values, the *only* neologism-proof equilibrium involves informative cheap talk.

Matthews (1987) shows how if there are equilibria of two kinds in a veto-threat game, then neither of them is neologism-proof. This is similar to the outcome in the Crawford–Sobel (1982) model: as I now show, their quadratic example has *no* neologism-proof equilibrium unless the uncommunicative equilibrium is the unique perfect Bayesian equilibrium.

In the Crawford–Sobel example, $T = A = [0,1]$ and the prior is uniform. Pay-offs are $u^R(a,t) = -(a-t)^2$ and $u^S(a,t) = -(a-t-b)^2$, where $b > 0$ is a parameter measuring the degree of conflict of interest. Crawford and Sobel show that, if $N(b)$ is the largest integer N satisfying $2N(N-1)b < 1$, then there is one perfect Bayesian equilibrium outcome for each integer N between 1 and $N(b)$. The N-equilibrium is described by the partition

$$[0,1] = [x_0, x_1) \cup [x_1, x_2) \cup \cdots \cup [x_{N-1}, x_N],$$

where

$$x_i = i/N + 2bi(i-N) \quad (0 \le i \le N).$$

If $t \in [x_{i-1}, x_i)$ then S sends some message that tells R (in equilibrium) precisely that $t \in [x_{i-1}, x_i)$, and so R chooses the action $a_i = (x_{i-1} + x_i)/2$.

Since S always wants a to be higher than R wants ($b > 0$), one might expect that, if t is very close to 1, then S will try to reveal that fact. We therefore investigate whether there is a self-signalling neologism of the form $Y = (y, 1]$, where $y > x_{N-1}$.

For Y to be self-signalling, it is necessary and, for $y \in (x_{N-1}, 1)$, sufficient, that when $t = y$, S is indifferent between using Y (and thus inducing the action $a(Y) = (1+y)/2$) and using his equilibrium message, thus inducing the action $a_N = (1 + x_{N-1})/2$. This requires the marginal condition that

$$(1+y)/2 - (y+b) = (y+b) - (1+x_{N-1})/2,$$

whence $3y = 3 - 1/N - 2b(N+1)$. If this equation gives a value of y in the range $(x_{N-1}, 1)$ then we have constructed a self-signalling neologism Y. It is immediate that $y < 1$. If $N \ge 2$ then indeed $y > x_{N-1}$. If $N = 1$ then $y > x_{N-1}$ if and only if $b < \frac{1}{2}$.

We conclude that if $b < \frac{1}{2}$—which, as Crawford and Sobel show, is a necessary condition for an equilibrium with communication—then there is no neologism-proof equilibrium. Thus in their quadratic example (though not, of course, in general) there is no neologism-proof communication. This result is certainly disturbing, and suggests that neologism-proofness is perhaps too strong a condition, although as I argued in §9 above, the absence of neologism-proof equilibrium need not be the end of analysis.

12. Conclusion

In cheap-talk games, standard refinements of perfect Bayesian equilibrium do not eliminate any equilibria. To eliminate unreasonable equilibria, we considered out-of-equilibrium messages with focal meanings: the literal meanings of unexpected messages (neologisms) in a natural language. This approach separates the problem of comprehension from the problem of credibility. In some equilibria, some neologisms are credible, and if (as seems reasonable) they would be believed, this eliminates those equilibria. We say that the equilibria are not neologism-proof.

In some cases, indeed, no equilibria remain. We could 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.

Games should be taken in context, especially when analysing the effects of communication. Language that could not survive in equilibrium if the world were nothing but a particular game, can nevertheless affect the outcome of that game.

References

Banks, J., and Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, **55**, 647–62.

Cho, I-K., and Kreps, D. (1987). Signaling games and stable equilibria. *Quarterly Journal of Economics*, **102**, 179–221.

Cramton, P. (1984). The role of time and information in bargaining. *Review of Economic Studies*, **167**, 579–94.

Crawford, V., and Sobel, J. (1982). Strategic information transmission. *Econometrica*, **50**, 1431–51.

Dawkins, R., and Krebs, J. (1979). Arms races between and within species. *Proceedings of the Royal Society of London, Series B*, **205**, 489–511.

Farrell, J. (1985). Credible neologisms in games of communication. Working paper 386. Massachusetts Institute of Technology.

Farrell, J., and Gibbons, R. (1988*a*). Cheap talk can matter in bargaining. *Journal of Economic Theory*. Forthcoming.

Farrell, J., and Gibbons, R. (1988*b*). Cheap talk, neologisms, and bargaining. Mimeograph. University of California–Berkeley, and Massachusetts Institute of Technology.

Gertner, R., Gibbons, R., and Scharfstein, D. (1988). Simultaneous signalling to the capital and product markets. *Rand Journal of Economics*, **19**, 173–90.

Green, J., and Stokey, N. (1980). A two-person game of information transmission. Mimeograph. Harvard University.

Grossman, S. and Perry, M. (1986*a*). Sequential bargaining under asymmetric information. *Journal of Economic Theory*, **39**, 120–54.

Grossman, S. and Perry, M. (1986*b*). Perfect sequential equilibrium. *Journal of Economic Theory*, **39**, 97–119.

Kohlberg, E., and Mertens, J. F. (1986). On the strategic stability of equilibria. *Econometrica*, **54**, 1003–38.

Lewis, D. (1969). *Convention*. Harvard University Press.

McLennan, A. (1985). Justifiable beliefs in sequential equilibrium. *Econometrica*, **53**, 889–904.

Matthews, S. (1987). Veto threats: rhetoric in a bargaining game. CARESS working paper 87–06. University of Pennsylvania.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge University Press, London.

Milgrom, P., and Roberts, J. (1982). Limit pricing and entry under incomplete information. *Econometrica*, **50**, 443–59.

Myerson, R. (1983). Mechanism design by an informed principal. *Econometrica*, **51**, 1767–98.

Palfrey, T., and Rosenthal, H. (1988). Testing for cheap talk effects in a public goods game with replay. Mimeograph. California Institute of Technology.

Riley, J. (1979). Informational equilibrium. *Econometrica*, **47**, 331–59.

Rubenstein, A. (1985). A bargaining model with incomplete information about time preferences. *Econometrica*, **53**, 1151–72.

Schelling, T. (1960). *The strategy of conflict.* Harvard University Press.

Selten, R. (1975). A reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, **4**, 25–55.

Spence, M. (1974). *Market signaling.* Harvard University Press.

Wittgenstein, L. (1958). *Philosophical investigations.* Blackwell, Oxford.