working paper
department
of economics

Communication In Games I:
Mechanism Design Without a Mediator

Joseph von R. Farrell

#334

December, 1983

massachusetts
institute of
technology
50 memorial drive
cambridge, mass. 02139

Communication In Games I:
Mechanism Design Without a Mediator

Joseph von R. Farrell

#334 December, 1983

D11

Communication In Games I: Mechanism Design Without a Mediator

Joseph von R. Farrell
Department of Economics
MIT, Cambridge, MA 02139

This is a working draft and incomplete in a number of ways which will be evident to the reader. Comments are earnestly solicited. Some have already been received, and have been most helpful. I particularly wish to thank Vincent Crawford, Jerry Green, Bengt Holmstrom, Eric Maskin, Andrew McLennan, Paul Milgrom, Roger Myerson, and Joel Sobel.

Revised December, 1983.

I. Introduction

Cooperation and mutual advantage are often limited by the presence of private information. A classic example is adverse selection in insurance markets. When agents interact in circumstances in which there is private information, Bayesian equilibrium is usually taken as the solution concept. Such an approach takes as fixed the structure of private information; but in fact it is clear that agents will sometimes want to give away their secrets and be able credibly to do so.

For example, in the following "cooperation game" (Farrell, 1982) the only Bayesian equilibrium[1] involves no cooperation, and gives payoffs of zero for the two players, whatever the actual levels of the "cooperation costs" y_i, which may in fact be close to zero.

		II's move	
		C[ooperate]	N[ot cooperate]
I's move	C	$(1-y_1, 1-y_2)$	$(-y_1,0)$
	N	$(0, -y_2)$	$(0,0)$

Here y_i is information private to player i; y_1 and y_2 are independently uniformly distributed on (0, 1.000001). (The result is more general: see Farrell, (1982).) Yet we would not expect such a poor outcome from two rational people involved in this game, unless we forbade all communication between them. Even quite simple communication can lead to a better outcome, in this game.[2] In this paper I examine the extent and the effects of communication (without commitment) in general games.

A related problem is known as "mechanism design." A mechanism is a function from $T = T_1 \times \ldots \times T_n$, the space of full descriptions of private information, to distributions of outcomes. The standard mechanism design problem has private information but no private actions. A central authority, either disinterested or equipped with powers of commitment not given to the agents themselves, can commit himself to a rule, describing the outcome as a function of messages he receives. Agents may of course be induced to lie if certain rules are chosen. A rule, and a Bayesian equilibrium in the game induced among the agents, determine a mechanism. The problem is: what mechanisms can be implemented in this way? The basic result is the revelation principle: every feasible mechanism is equivalent to a full-revelation mechanism, i.e. one in which each agent voluntarily and truthfully reveals his entire private information to the central authority. A variation on this literature (see e.g. Laffont and Maskin (1982); Myerson (1983)) discusses the case in which the center is not an omnipotent authority but only a mediator who controls the communication, while agents have private actions: a version of the revelation principle applies there too. It says that without loss of generality we can further assume that each agent, told only his proper action, will obey. (Actually, the distinction is not as clear as might be thought, since most power comes from control of communication: a dictator seldom <u>does</u> what he wants done, but rather sends messages which make it in the perceived interests of others to implement those decisions.)

What will happen if there is no such discreet and disinterested mediator? Often there is not, either because everyone available is involved in the outcome, or because the cost of finding and checking a discreet and disinterested person is excessive for the problem at hand. In fact, one way to view our problem is as part of the question, "When will agents appoint a mediator; and how much power will they give him?" (It should be noted that the "mediator" need not be a person: one could use a properly programmed computer. But this solution is seldom adopted.)

In Section II, I introduce a notion of mechanism feasibility without a mediator, appropriate for the case in which agents can together commit themselves to a format for communication. It is unlike the standard problem with a mediator, because players are cautious about how much they say, since it becomes public knowledge: there is no revelation principle, because the interested agents cannot commit themselves to ignore information. I show by example that this makes a real difference to the class of feasible mechanisms. In other words, the mediator's function cannot always be decentralized, even if that function is only collection and dissemination of information, not decision. Another way of viewing the problem is to describe the outcomes of G which become equilibria when arbitrary payoff-irrelevant choices are added.

In Section III, I address the more difficult case in which agents cannot commit themselves to limiting communication. This requires eliminating some perfect Bayesian equilibria as implausible; I offer what I believe is a new criterion which seems to conform well to intuition about this type of implausibility. (The criterion is also useful in examining the relevance of communication for disequilibrium, work I am pursuing in a related paper.) This leads to a smaller class of equilibria, which I call "language-proof," and a corresponding class of feasible equilibrium mechanisms, which I call "variable-structure communication-feasible" (VSCF). Some examples are presented.

The purpose of this work is to understand the effects of decentralized communication, and the value of being able to centralize it. We already have fairly well-developed notions of what can be expected to happen without communication, and of what is feasible with an ideal mediator. In this paper I address the intermediate case, in which there can be communication among players (who share a language), but there is no mediator.

II. Fixed-Structure Communication-Feasible (FSCF) Mechanisms

A group of agents, numbered 1, 2,..., n, are to play a game G with incomplete information: agent i has private information $t_i \varepsilon T_i$. There is no mediator to help them make joint use of their information; however, they are alone together in a room for some while before G begins - and, if G is sequential, between moves - and they can talk. What will be the outcome?

As usual in this literature, I look for equilibria,[3] though of course the precise game of which we look for equilibria is yet to be specified. I will also assume that anything said by one player is heard by all. This is restrictive[4] if n > 2.

Consider the process of communicating, and playing G, as an "extended game" or "communication version of G," which we can call G'. We look for perfect Bayesian equilibria in G', and call the resulting mechanisms "fixed-structure communication-feasible" (FSCF) mechanisms in G. There will of course be many possibilities for G', some differing only inessentially, and others perhaps leading to quite different mechanisms. For the purposes of this section, we include them all. This is intended to capture the idea that, while the structure of G itself may be exogenous, we have no basis yet for considering one possible communication version G' rather than another. Thus our definition is intended to capture all outcomes that are self-enforcing with some communication structure. (In Section III, I distinguish some implausible structures.) I now give the formal definition.

A game G' is a communication version of G if it is a sequential game of incomplete information, with the same spaces T_i of private information, and has the following properties:

(i) The possible histories (paths) in G' are the possible histories (paths) in G, but with (perhaps) some extra moves, which I call "message moves," inserted.

(ii) The payoffs from a history in G' are just the payoffs obtained by "condensing" to a G-history by ignoring the message-moves, and combining the G-moves with the private information as if the G-history had happened in G. In other words, the message-moves are "payoff-irrelevant." (However, this does not mean that their presence does not affect payoffs!)

(iii) There is perfect information concerning the messages sent in message-moves. In other words, when player i sends message m_i at a given point, that fact then becomes common knowledge.

In a communication version of G' of G, each vector of decision rules, and in particular each perfect Bayesian equilibrium, determines a mechanism in G, that is, a function $m\colon T \to \Delta(A)$, where $\Delta(A)$ represents the space of distributions on $A = A_1 \times \ldots \times A_n$, the outcome space. A mechanism m is called

<u>fixed-structure communication-feasible</u> (FSCF) if it arises from a perfect Bayesian equilibrium in some communication version G' of G.

We observe:

(1) Every perfect Bayesian equilibrium mechanism of G is also FSCF. This is immediate: set G' = G.

(2) In general, the class of FSCF mechanisms is strictly larger than the class of perfect Bayesian equilibrium mechanisms. An example is provided by the cooperation game (page 2). It is easy to see that if each player has a chance to communicate one binary message, the following mechanism will arise in an equilibrium:

(C,C) if both the y_i are less than or equal to 1.

(N,N) otherwise.

(3) Every FSCF mechanism is feasible using communication via a mediator, i.e. incentive-compatible in the sense of Myerson (1983). This is immediate, since the mediator can promise to mimic any FSCF mechanism. An example of a Myerson incentive-compatible mechanism which is not FSCF is given below. With the kind of public communication I consider, only certain kinds of information about $T = T_1 \times \ldots \times T_n$ can be communicated: I call this "announceable" information. In the example below, the desired mechanism depends on certain non-announceable information being communicated.

We can view the difference between these two classes of mechansims as representing the value (to the players) of having available a suitable mediator, if without a mediator they can agree on and commit to a communication structure.

Announceable Information

Even if we ignore incentive constraints, not all possible systems of partial or total public revelation of $\underline{t} \in T$ can be generated by sequences of public announcements. The essential point is this: It is impossible to have the extent of i's revelation depend on what j reveals, and the extent of j's revelation depend on what i reveals. This can make certain desirable outcomes infeasible, as in the case of the "strength and weakness" example below. It is a problem that a mediator can overcome, by having both i and j reveal their information privately to him. In this sub-section, I describe the class of revelations, or information systems, which can emerge from public communication. This should prove a useful step towards classifying the FSCF mechanisms, although the interrelationship with incentive problems is likely to be difficult. It will also address more directly the class of problems in which (i) lying is detectable and adequately punishable ex-post, but (ii) private actions cannot be controlled, so that it may be desirable to arrange for less than complete revelation. Then we can predict the outcome from each posterior on $\underline{t}$, and the class of feasible distributions of posteriors (if below) then gives the class of feasible distributions of outcomes.

Describing Revelation Systems

For our present purpose, we ignore the Cartesian product structure of T ($T = T_1 x \ldots xT_n$), and simply list the elements: $T = \{t^1, \ldots, t^n\}$. We assume that, before receiving private information, everyone has the same prior $\underline{\pi} = (\pi_1, \ldots, \pi_n)$ on T; that is,

π_i = prior probability that t^i is the true state.

Revelations will update these beliefs. We model this as follows. Define the simplex S as:

$$S = \{(p_1,\ldots,p_n): \ p_i \geq 0; \ \Sigma\, p_i = 1\}.$$

Consider probability distributions on S. The prior $\underline{\pi}$ is represented by the distribution all of whose mass is concentrated at the point $\underline{\pi} \in S$. Any revelation corresponds to a mean-preserving spread on S. For instance, suppose that with probability $p(m_j|t^i)$, if t^i is the true state, message j is the result. $(i=1,\ldots,N; j=1,\ldots,m)$. The the posterior probability of t^i, if m_j is heard, is

$$p[t^i|m_j] = \frac{p[m_j|t^i]p[t^i]}{p[m_j]}$$

$$= \frac{p(m_j|t^i)\pi_i}{\sum_{i=1}^{N} p(m_j|t^i)\pi_i}$$

and we represent this message system by the following distribution on S: A probability mas $p[m_j] = \sum_{i=1}^{N} p(m_j|t^i)\pi_i$ is attached to the point $\underline{p}(m_j) = (p[t^i|m_j], p[t^2|m_j],\ldots,p[t^n|m_j])$. It is easy to check that this distribution is a mean-perserving spread of $\underline{\pi}$, and that (using Bayes' rule) one can recover the initial data $p(m_j|t^i)$ from the distribution.

Proposition 1. Assume that the agents' information jointly determines $t \in T$, so that, if all were revealed, no uncertainty about t would remain. Then (incentive constraints aside) a mediator can construct a (partial) revelation mechanism for each distribution on S which is a mean-preserving spread of $\underline{\pi}$.

Proof: I restrict the proof to distributions with finite support $(\underline{p}^1,\ldots,p^m) \leq S$. It would not be hard to extend it to general distributions. To achieve the distribution with weight q_j on $\underline{p}^j$ (i=1,...m), the mediator (having collected all private information, and therefore knowing t^i), announces message "j" with probability $p[j|t^i]$, where

$$p[j|t^i] = \frac{p[t^i|j]p[j]}{p[t^i]} = \frac{p_i^j q_j}{\pi_i}$$

By construction, this will indeed generate the right posteriors. We need only check that $\sum_j p[j|t^i] = 1$ for each i:

$$\sum_j \frac{p_i^j q_j}{\pi_i} = \frac{1}{\pi_i} \sum\nolimits_j p_i^j q_j = 1$$

but this follows since the mean of the assigned distribution is $\underline{\pi}$.

Q.E.D.

Now, in general, it will *not* be the case that every distribution with mean $\underline{\pi}$ can arise by public announcements. To understand this, recall that no one player has all the information, in general. (If one does, then the problem disappears.) Thus, for each player, h, there are states t^i, t^j which are indistinguishable. Therefore no announcement by h can change the relative probabilities p^i/p^j.

Definition. Let d be a distribution on S, attaching weight $d(\underline{p})$ to each point of a finite support $\{\underline{p}^1,\ldots,\underline{p}^m\}$. Let h represent an agent ($h=1,2,\ldots,n$). A distribution d' is an h-stretch of d if it can be obtained from d by the following changes:

For each point $\underline{p}^k$, take some of the weight $d(\underline{p}^k)$ from $\underline{p}^k$, and distribute it within the set $D^h(\underline{p}^k) = \{\underline{p} \in S$: if i and j are indistinguishable to h, then $p_i/p_j = p_i^k/p_j^k\}$, in such a way as to leave the mean at $\underline{p}^k$.

The idea of this is that h can make different announcements depending on what has already been announced (i.e. the different $\underline{p}^k$'s), but this announcement can depend only on his own information. (At the cost of some slight increase in complexity, one could consider distributions with general supports.) There is perhaps a slight loss of generality from the fact that we allow h's announcement to depend only on beliefs just before his announcement, not on how those beliefs were arrived at. I suspect this will not matter much.

Definition: A distribution d on S is announceable if it can be obtained, starting from $\underline{\pi}$, by a sequence of stretches.

Proposition: Not every distribution on S with mean $\underline{\pi}$ is announceable.

Proof: Rather than constructing the simplest counterexample, we discuss a more general class. Suppose there are n players, and T can be written as $T_1 x \ldots x T_n$, where h knows $t_h \varepsilon T_h$, and where t_h is independent of $t_k (k \neq h)$. Suppose T_h contains r members (it would be easy to allow $|T_h| \neq |T_k|$, but still not worth the effort). Write $p(i_1, i_2, \ldots, i_n)$ for the probability that $t_1 = i_1$, $t_2 = i_2, \ldots,$ $t_n = i_n$, where $1 \leq i_1, \ldots, i_n \leq r$. Then by independence, if p is the prior,

$$p(i_1, \ldots, i_n) p(i_1', i_2', \ldots, i_n') = p(j_1, \ldots, j_n) p(j_1', \ldots, j_n') \tag{1}$$

where, for each h, j_h and j_h' are i_h and i_h' but not necessarily respectively. Moreover, this relationship is preserved by every stretch, that is, by every individual revelation. To see this, fix h, and suppose (without loss of generality) that $j_h = i_h$, $j_h' = i_h'$. Starting with a p satisfying (1), suppose p' is given weight in an h-stretch. Then we know, since $(i_1, \ldots, i_n)$ and $(j_1, \ldots, j_n)$ are indistinguishable to h, that

$$\frac{p'(i_1, \ldots, i_n)}{p'(j_1, \ldots, j_n)} = \frac{p(i_1, \ldots, i_n)}{p(j_1, \ldots, j_n)} \tag{2}$$

and likewise that

$$\frac{p'(i_1', \ldots, i_n')}{p'(j_1', \ldots, j_n')} = \frac{p(i_1', \ldots, i_n')}{p(j_1', \ldots, j_n')} . \tag{3}$$

Since we can express (1) as

$$\frac{p(i_1, \ldots, i_n)}{p(j_1, \ldots, j_n)} = \frac{p(j_1', \ldots, j_n')}{p(i_1', \ldots, i_n')} \tag{4}$$

it follows that the same relationship (4) must hold for $\underline{p}'$. Thus, (1) holds for every point $\underline{p} \in S$ which is given weight in an announceable distribution. This proves the Proposition; it also tells us that announceable distributions are supported on the subset of S on which all relationships of the form (1) hold.

For example, consider the case $n = 2$, $r = 2$, which we will discuss below in the "strength & weakness game." Then we have just one constraint, namely

$$P(1,1)\ p(2,2) = p(1,2)\ p(2,1)$$

or, writing W for 1 ("weak") and S for 2 ("strong"),

$$p(w,w)\ p(s,s) = p(w,s)\ p(s,w) \qquad (5)$$

Revealing "whether or not both are weak" involves putting weight on a point with

$$p(w,w) = 0 \qquad (6)$$

$$p(w,s) = p(s,w) > 0$$

which does not satisfy (5). Thus, that information can not be revealed, even approximately, with any sequence of (stochastic) announcements. It would be desirable to characterize the announceable distributions on S. At present, we only have two necessary conditions:

(i) the mean of the distribution must be $\underline{\pi}$.

(ii) the support must be contained in the smallest set which contains $\underline{\pi}$ and which contains $D^h(\underline{p})$ whenever it contains $\underline{p} \in S$, and for all h.

It seems clear that these conditions are not sufficient, however.

Application: Bargaining, Two-Sided Uncertainty.

In bargaining, each of the two players has private information about his own value of the item they are bargaining over. One question we might ask is: Would it be possible for the players first to find out, jointly, whether or not there are potential gains from trade? If player 1 is initially the owner of the item, and has value v_1 for it, while player 2 has value v_2, we ask whether it can be made common knowledge whether or not $v_2 > v_1$. This amounts to announcing the distribution:

$$\text{Prob}[v_2 > v_1] \text{ at the point } E((v_1,v_2)|v_2 > v_1) \in S$$

$$\text{Prob}[v_2 \leq v_1] \text{ at the point } E((v_1,v_2)|v_2 \leq v_1) \in S.$$

We ask whether this distribution is announceable. Suppose we can choose v_{11}, v_{12} (possible values of v_1) and v_{21}, v_{22} (possible values of v_2) such that all four combinations are possible, and such that

$$v_{21} > v_{11}$$

$$v_{22} < v_{11}$$

$$v_{21} < v_{12}$$

$$v_{22} < v_{12}$$

Then we are asking for an information system which puts positive weight on a posterior with positive weight on (v_{11},v_{21}) and zero weight on each of the other three combinations. But, with independent values, this is not announceable. The desirable information would, however, be extractable with a mediator (truthtelling would be an equilibrium), and so this shows the problems generated by decentralized communication.

Announcability and Incentive-Compatibility

So far, in discussing announceability, I have ignored incentive problems. In order to deal with games of public communication, however, even in the "classical" case in which there are no private actions, we must consider the incentives to lie.

Consider a mechanism which is incentive-compatible in the usual sense with a mediator, so that no player wishes to lie to the mediator. Suppose also that the information needed to implement the mechanism is announceable. Does it follow that the mechanism can be implemented using only public communication? In general the answer is no. The reason is that, in the usual incentive-compatibility calculation, incentives to lie are calculated (for each player) on the basis of his own private information only. However, in using public communication to generate an announceable revelation, some players will hear some information about others before making their own revelations. Thus they know more when considering whether to lie, and this means that a stronger incentive-compatibility criterion is needed. If truthtelling is a dominant strategy, that is sufficient, but that is relatively rare. Likewise, if the revelation is announceable simultaneously, the standard condition is sufficient.

The Strength/weakness Game: An example showing positive value of a mediator.

In this game, there is useful information which cannot be conveyed between the two players because it is not announceable. Moreover, no announceable refinement of the desired information can be communicated, because of incentive-compatibility problems. Thus the only communication-feasible mechanisms involve not communicating that information (in this example, the only information).

There are two players, I and II. Each observes whether he is weak (W) or strong (S). The types are independent, and the probability of S exceeds one half. Each player has two moves, called X and Y. The payoffs are as follows (I chooses the row, II the column):

Both Strong:

	X	Y
X	(1,1)	(0,0)
Y	(0,0)	(0,0)

Both Weak:

	X	Y
X	(0,0)	(0,0)
Y	(0,0)	(1,1)

I strong, II weak:

	X	Y
X	(0,0)	(-2,0)
Y	(1,-2)	(-2,-3)

I weak, II strong:

	X	Y
X	(0,0)	(-2,1)
Y	(0,-2)	(-3,-2)

Consider the following desirable mechanism: Each player plays X, unless both players are weak, in which case each plays Y. This is incentive-compatible using a mediator: Each player reveals his type to the mediator, who announces "play Y" only if he heard two "W" messages. If no such announcement is heard, each player is content to play X; if it is, each wishes to play Y, provided he believes the other player told the truth to the mediator. It is also straightforward to show that truthtelling is a Bayesian equilibrium,[5] given the mediator's rule.

However, the information "whether or not both are weak" is not an announceable partition, and so cannot be communicated without a mediator. Moreover, the players cannot use a refinement: a weak player will not be prepared to announce his weakness first to the other, since the potential loss if the other is strong outweighs the potential gain if the other is also weak. The point is that, if II believes I is weak, he will play Y. This is desirable for I if and only if in fact both players are weak. However, it is impossible, without a mediator, for a weak I to make his revelation conditional on II also being weak. Thus, in this game, the opportunity for communication will not be taken unless there is a mediator available.

III. Language-proof Equilibria and Variable-Structure Communication-Feasible (VSCF) Mechanisms

In the example presented above, a mediator can facilitate communication which will not occur without him. However, in other cases too much communication may be at once harmful and tempting, and a mediator's role may - paradoxically - be to prevent inappropriate communication. In this section, I offer a new theory of when (absent a mediator) unauthorized communication may be expected. This leads to the idea of a language-proof equilibrium, and a VSCF mechanism: one in which no unauthorized communication will happen. Intuitively, an equilibrium which is not language-proof is only self-enforcing if agents are somehow physically unable to make claims which will be believed if made, and which are worth making if true. Unless there are ways of committing agents not to engage in such unauthorized communication, an equilibrium (in the perfect Bayesian sense) which is not language-proof is not self-enforcing, and therefore does not properly implement the corresponding outcome rule $r: T \to \Delta(A)$. Thus, we have a class of rules called variable-structure communication-feasible (VSCF), which can be implemented by language-proof equilibria. Every VSCF rule is FSCF, but the converse is by no means true. Allowing for the possibility of unauthorized communication reduces the class of self-enforcing communication structures, and hence the class of feasible rules.

Thus, the difference between perfect Bayesian equilibrium and language-proof equilibrium, or between FSCF and VSCF, is that, in the former, it is assumed that agents can commit themselves to a format of communication - in particular, sometimes to remaining silent. When such commitment is impossible, we are led to the class of language-proof equilibria and the corresponding class of VSCF rules.

When Communication May be Inappropriate, But Tempting

Consider a game with two players, I and II, and let I have private information. Even if I's revelation of his information is voluntary and desirable for him, it may harm II. Of course, II likes having information, but he may dislike the effects of I's knowing he (II) has the information. Thus, II can be better-off if communication is infeasible.

In fact, I and II may both be better-off. We can construct examples in which, if I discovers a "good" state, he wants to tell II, but, if he discovers a "bad" state, he wants II to remain uncertain. If it is common knowledge that communication is feasible, however, II will not remain uncertain if he hears no "good" message; he will conclude that the state is "bad." In some cases, this can mean that both players, ex-ante, would prefer that there be no communication channel. (See Appendix.) However, they may be unable to arrange this.

Thus, it can happen that desirable mechanisms are ruled out by agents' inability to commit themselves to not communicating, or to restricted forms of communication. In this section, I discuss when and why a perfect Bayesian equilibrium may be vulnerable to unauthorized communication. To discuss this, it is not enough to ask whether the equilibrium is robust to the addition of more communication channels, as one might think. The reason is that every equilibrium is robust in that sense, but in an unconvincing way.

If we add message channels, there is always an equilibrium in which the sender fills the channels with messages indistinguishable from those he might use to communicate, but actually uncorrelated with any relevant information, and for those who receive these meaningless "messages" to ignore them.

(Technically, let G' be a communication version of a game G, and let E be any perfect Bayesian equilibrium of G'. Now let G" be another communication version of G, but with more message opportunities. (G" is a communication version of G'.) Then there is a perfect Bayesian equilibrium E' in G", in which all the additional channels are filled with "noise" as above, and which gives the same results as E.)

These equilibria, which might be called "noisy" equilibria (because the communication channel is effectively destroyed by the noise), are widely regarded as implausible. Kalai and Samet (1983) impugn the "persistence" of a noisy equilibrium. The view I take here is intuitively based on the idea that message-sending is very slightly costly,[7] so that an agent wishing to convey no information will simply remain silent. (As Lehrer once said, "If people can't communicate, the very least they should do is to shut up.")

This brings us to the second way in which a communication channel can be neutralized in perfect equilibrium. It is always an equilibrium for no message to be sent, if the response to any message would be to ignore it - in other words, not to understand it. If one believes that meaning is generated only in an equilibrium, it is reasonable to believe also that any new "unauthorized" messages will not be understood, since they do not occur (and thus acquire meaning) in equilibrium. Thus the standard approach is internally consistent, and is appropriate when there is no world outside the game. In this paper I take a different approach, and assume that there is a sufficiently rich and flexible language already in place. This makes it natural to place certain restrictions on interpretation of messages. In particular, a message which satisfies a certain condition (<u>self-signaling</u>)

is plausible, and I assume that such a message will be not only understood but believed. In some perfect Bayesian equilibria, there are self-signaling messages, which disrupt the equilibrium. I say those equilibria are not *language-proof*, and therefore not self-enforcing.

Informally, I assume[8] that the messages to be considered are messages of the form "$t_i \varepsilon S$," where S is a non-empty subset[9] of i's information set T_i. I assume that, for every such S, there is a message m(S) which is not sullied by being used with noise, or as a lie, in the equilibrium; and that it is common knowledge that the dictionary meaning of m(S) is that $t_i \varepsilon S$. This "dictionary meaning" is understood to be the meaning of m(S) if there were no lies. Since lying is possible, there is no guarantee that m(S) would be used to mean S; however, this is a focal meaning, from which worrying about lies can begin.

Assumption: If an equilibrium is such, and $S \leq T_i$ is such, that i prefers to have other players believe S (rather than equilibrium beliefs) precisely when $t_i \varepsilon S$, then S is *self-signaling*, and m(S) is interpreted to mean $t_i \varepsilon S$, and this is common knowledge.

Thus an "equilibrium" in which S is not meant to be communicated, but in which S is self-signaling, will not survive. If there is no such S, I say the equilibrium is *language-proof*. This means that, for every unauthorized message, there is some reason not to trust it.

In the formal definition below, I consider the simple case in which only player 1 can communicate, and only before G begins. (Later, I discuss extension.)

In this simplest case, we can "collapse" the problem to a payoff function $u_1(S_1,t_1)$, giving 1's (expected) payoff as a function of his true private information $t_1 \in T_1$, and of the subset $S_1 \leq T_1$ to which all other players' beliefs have been restricted by their interpretations of what player 1 has said. This function is supposed to be common knowledge; notice also that is is defined even when $t_1 \notin S_1$, since 1 could lie.

Take an non-empty $S_1 \leq T_1$, and define:[10, 11]

$S_1^* = \{t_1: u_1(S_1,t_1) > u_1(P(t_1),t_1)\}$, where $P(t_1)$ is the information conveyed about t_1 in equilibrium $P(\cdot)$.

If $S_1^* \leq S_1$, we say S_1 is *credible*. If $S_1^* \geq S_1$, we say S_1 is *unconcealed*. If $S_1^* = S_1$, we say S_1 is *self-signaling*. (Notice that it is precisely the change in beliefs itself that changes the payoff, not an exogenous cost.) The interpretation is that the message "$t_1 \in S_1$," which does not occur in equilibrium, will be sent precisely when it is true. To check whether a set S_1 is self-signaling requires only public information. I define an equilibrium $P(\cdot)$ to be *language-proof* if there exists no self-signaling S_1. A mechanism in a game G is VSCF if it can be implemented by a language-proof equilibrium in some communication version G'.

This defines what is self-enforcing when it is common knowledge that players are rational, have a language in common, and are used to people speaking the truth except when it would benefit them to lie. Allocations we normally consider equilibra may not be self-enforcing in this sense.[12] I now give some applications of the definition.

Application 1: Uniform Preferences.

We say 1's preferences are uniform if, for all t_1, $t_1' \varepsilon T_1$ and for all non-empty S_1, $S_1' \leq T_1$,

$$u_1(S_1,t_1) > u_1(S_1',t_1)$$

implies

$$u_1(S_1,t_1') > u_1(S_1',t_1').$$

One might expect that, if preferences are uniform, every equilibrium is language-proof: intuitively, it would seem, nothing can be signaled. This is not quite true, as shown by the following example, in which player 1 observes the state (A or B), communication occurs, and player 2 chooses move X, Y or Z: The payoffs are:

State A:

Move	1's payoff	2' payoff
X	0	3
Y	0	0
Z	1	2

State B:

X	0	0
Y	0	3
Z	1	2

In this game, there is an equilibrium in which 1 tells 2 the true state, 2 chooses X in state A, Y in state B, and in every state 1 gets 0 and 2 gets 3. This equilibrium is not language-proof: the set {A,B} ("I refuse to tell you the state") is self-signaling. The unique language-proof equilibrium involves no communication. (Compare this with the example in the Appendix, in which only one type prefers vaguer beliefs, and the unique language-proof equilibrium involves full communication.)

Notice how, in this example, there is an implicit question of commitment. The usual equilibrium story allows 2 to be "committed" to (say) treating every unauthorized message as a statement that state A has occurred. Although in equilibrium this commitment is never tested, there is still a question of its credibility. 2 has to claim convincingly that he will not understand if 1 announces "I won't tell you the state. Notice that this is a good plan for me whichever state has happened." In requiring Language-proofness, we do not allow 2 to convince 1 of such an inability.

Notice also that this example shows that a full-revelation equilibrium is not, as might be thought, necessarily language-proof.

Returning to the general case of uniform preferences, it is true that the only possible self-signaling set is T_1 itself. This shows that the no-communication equilibrium is language-proof, and also that any equilibrium in which there is at least one p_j which is (non-strictly) preferred to T_1, is language-proof.

Many models in the literature involve uniform preferences. Signaling models (Spence, Milgrom-Roberts, Kreps-Wilson-Milgrom-Roberts), adverse selection models (Akerlof, Wilson), and auction-sales models are examples. The lesson is that pay-off irrelevant ("costless") communication will not disrupt any equilibrium. If the relevant class of possible messages are such that, in some sense, indifference is not a real possibility (as in the examples just cited) then payoff-irrelevant communication will not occur in equilibrium, either.

Application 2 (Crawford-Sobel)

Crawford and Sobel (1982) consider a stripped-down communication-in-games model, in which one agent (R) has no private information, while the other (S) has no payoff-relevant choices. (Green and Stokey (1980) also considered such a model.) Crawford and Sobel show that there are a number of equilibria, differing in the size of the equilibrium partition of the state space $[0,1]$. I shall show that, at least in the example they work out (their Section 4), no equilibrium is language-proof.

In the notation introduced by Crawford and Sobel, the state m is distributed uniformly on $[0,1]$; the action y is chosen by R after hearing S's message, and the payoffs are, for some given $b > 0$,

$$U^S = -(y-(m+b))^2$$

$$U^R = -(y-m)^2.$$

An equilibrium consists of a partition of $[0,1]$ into intervals (because of concavity). It is completely determined by a_1, where $[0,a_1]$ is the first of those intervals. I shall show that, unless $a_1 \leq 4b$, the equilibrium is not language-proof. The condition $a_1 \leq 4b$ is satisfied by the most-informative equilibrium, and by no other.*

Take a Crawford-Sobel equilibrium, and ask whether there exists $S = [0,s)$ which is self-signaling (credible and unconcealed), where $0 < s < a_1$. If so, it must be that, at $m = s$, the value to S of persuading R that $m \in S$

*It is simple to show that, if a_1 and a_1' correspond to different equilibria then $|a_1 - a_1'| > 4b$.

is equal to the value of persuading him that $m \in [0,a_1]$. Simple calculatios shows that this requires $s = \frac{1}{3}(a_1 - 4b)$. Clearly, then, $a_1 > 4b$ if there is to be such a self-signaling S. However, it is also simple to calculate that, provided $a_1 > 4b$, $S = [0, \frac{1}{3}(a_1 - 4b))$ is indeed self-signaling. Thus we have shown that every Crawford-Sobel equilibrium with $a_1 > 4b$ fails to be language-proof.

Since the upper end of the interval is symmetric with the lower end except for the fact that $b > 0$, consideration of possible self-signaling sets of the form $S = (1-s,1]$ can be expected to give the result that there is always such a set. This is indeed the case.

To interpret this result in terms of evolution of language, suppose we begin at a Crawford-Sobel equilibrium. In evolutionary terms, every S is genetically programmed to send message i when $m \in [a_{i-1},a_i]$, and every R is genetically programmed to take action $y_i = \frac{1}{2}[a_{i-1} + a_i]$ on hearing message i; and every mutation on the part of R, or S if it does not involve new messages, is disadvantageous.

Now suppose that when $m \in [a_{N-1},a_N]$ $(a_N=1)$ and $m > s$, where $(1-s) = \frac{1}{3}((1-a_{N-1}) + 4b)$, S sends the message N, but also sends a further message "+". The "richness" assumption states that this is possible. Initially, this has no effect on S's expected payoff: the mutation is neither advantageous nor disadvantageous. However, suppose that various mutations of R now arise, with different conjectures about what should be inferred from the signal "N+". Those whose conjectures closely enough approximate "$m \in (s,1]$" will be favored. Moreover, as the proportion of those R's grows, the pressure on S's will now be strictly in favor of the original mutation. Thus, if that mutation managed to survive long enough for the R's to respond, it will prove favorable.

In the human-language interpretaiton, suppose S sends an unexpected message "I announce not only that $m \in [a_{N-1},1]$, but also that $m > s$." R can calculate that this is self-signaling, and this makes it plausible that he will believe it. Hence, provided S is capable of saying what I just described, the equilibrium will not survive.

A simpler example in which language-proof equilibrium does not exist is the following. Player 1 knows the state (A or B). Player 2 will take actions X, Y, or Z according as he believes the state is A, is B, or is uncertain. Payoffs to 1 are

	A	B
X	0	1
Y	1	2
Z	2	0

The no-revelation equilibrium is not language-proof ($\{B\}$ is self-signaling), while revelation is not an equilibrium.

Application 3: The Cooperation Game. Consider the cooperation game (see p. 2) when $F_2(\cdot)$ is degenerate, so that y_2 is known and is less than 1. Consider an uncommunicative equilibrium. If $F_1(1) < y_2$, the only such equilibrium is no cooperation. Thus, $u_1(T_1, y_1) = 0$ for all y_1. However, if player 1 credibly announces $S_1 = (0,1)$, he gets more than 0 if the outcome (given common knowledge that $y_1 < 1$, $y_2 < 1$) is cooperation. If we allow player 1's intended move to be communicated along with (or as part of) his private information, the outcome *will* be cooperation if $y_1 < 1$. Hence, the uncommunicative equilibrium is *not* language-proof. This conforms to our intuition that, in this game, an opportunity for communication will not be wasted by going to the "noisy" equilibrium. I formalized this, using an "infinitesimal cost of communication," in Farrell (1982).

A Refomulation: Coalition-Formation.

Some insight into the nature of language-proof equilibrium, and the existence problem, may be provided by the following interpretation as a coalition-formation problem.

T represents a set of players t, each maximizing his own utility. In an equilibrium, t will be in just one coalition, $P(t)$, and his utility is $u(P(t),t)$. The function $u(S,t)$, where $t \in T$ and $\phi \neq S \leq T$, is exogenously given, and defined whether or not $t \in S$.

A *limited-language equilibrium* is a partition of T into coalitions P, such that, for all t, all P,

$$u(P(t),t) \geq u(P,t),$$

where $t \in P(t)$. Notice the difference[13] between this and

$$u(P(t),t) \geq u(P \cup \{t\},t)$$

which would represent a straightforward condition "t does not prefer to join some other coalition P." Here, instead, we can think of it as follows: The treatment of members of a coalition depends on the set of its official members. In equilibrium, all members must be official, but a player deviating becomes an unofficial member. The function u(P,t), where t ε P, expresses the payoff to an unofficial member.

Now consider a different kind of deviation: forming a new coalition S. If a new coalition is formed, there has to be an announcement to describe its membership. Moreover, that announcement must be credible: that is, given that the announcement describes the set of official members, a referee checks that (i) all the official members really wish to join, and (ii) nobody will wish to join unofficially. If the proposed coalition does not meet these requirements, it is not allowed to form; if it does, then it is allowed.

An equilibrium of this game is then a language-proof equilibrium of the ccrresponding communication game.

This coalition-formation view is related to that of Myerson (1983b), Section 7, but my concept of blocking by formation of a self-signaling coalition does not fit into Myerson's framework, because his Axiom 1 (Domination) fails to hold. The reason is this: A proposed equilibrium E (in Myerson's terms, a mechanism μ) may be blocked by a self-signaling set S. If E' is another equilibrium, worse for every type of player 1 than E, then every t ε S would prefer to defect to S, but so might some other types. There is no guarantee in general that there will be any self-signaling S' which blocks E'.

Generalizing the Definition

So far, I have only defined when an equilibrium is language-proof in the case where just one player can communicate. It is not hard to generalize this. First, consider the last point at which effective communication is possible. The definition given above for player 1 must hold for each player with a physical ability to communicate. Since we are considering the last point at which communication can happpen, reaction to unauthorized communication cannot involve further communication. Hence it is relatively straightforward to define the payoffs.

Now, consider the time next before that. (If time is not discrete, problems may arise.) Any communication, authorized or not, leads to a subgame. We require that the subgame be assigned a language-proof equilibrium. Now evaluate any proposed communication using that behavior in the subgame, and require that the second-level-up subgame be conducted in a language-proof way.

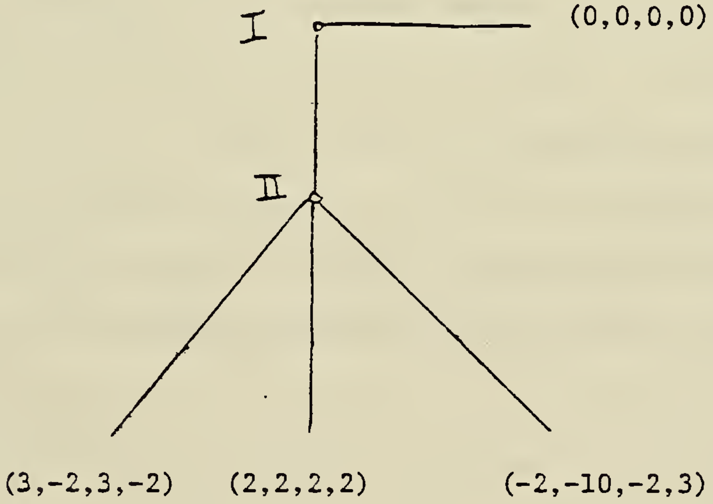
Thus, language-proofness is defined recursively starting from the end. One difficulty with this is that, if a subgame has no language-proof equilibrium, we do not know what payoffs to assing to reaching it. This is a problem even if we do not reach it in a proposed equilibirum. The question is, how do players, contemplating an unauthorized communication which will lead to chaos (no language-proof equilibrium), evaluate that option? In some fortunate cases we could use bounds on the payoffs, but this will not generally suffice.

Appendix

Example in which too much information is revealed.

We construct* an example with the simplest possible private information structure: one player (I) has one binary piece of private information: he knows whether state 1 or state 2 has occurred. We need a game in which player II's action is different according to whether he knows state 1, knowe state 2, or knows nothing; moreover, I must behave differently (at least in one state) when he knows II knows the state than when he (I) knows II does not. Thus II must have at least three moves, and I must have at least two.

Consider the following sequential game. The payoff vectors are written thus: (Payoff to I in state 1, to I in state 2, to II in state 1, to II in state 2).



*This example is based on one produced during a conversation with A. McLennan, to whom I am grateful.

Without communication, a feasible mechanism is for I always to play "down" and for II always to play "center." This gives each player a payoff of 2.

However, suppose there is time and language available for I to communicate, even though he has not intended to do so. The message "state 1 has occurred" is self-signaling:

- <u>credible</u>, since if II believes it, he will play left if I plays down, and this would be bad for I_2.
- <u>unconcealed</u> since I_1 can get 3 by persuading II of it, when it is true.

Thus this equilibrium is not language-proof. In fact, the only language-proof equilibrium involves communication of the state, and outcomes:

(down; left) in state 1

(across; right) in state 2.

This gives ex-ante payoffs (3/2, 3/2), strictly less than the payoffs from the FSCF (no communication) mechanism mentioned above.

FOOTNOTES

[1] In fact, this is the only rationalizable outcome (for definition, see Bernheim, 1982) whenever y_1 and y_2 have distribution functions F_1, F_2 on $(0,\infty)$ such that $F_1(F_2(x)) < x$ for all strictly positive x. See Farrell (1982).

[2] With one opportunity for player I to communicate a binary message, the following allocation is an equilibrium mechanism:

(C,C) if both $y_1 \leq F_2(1)$ and $y_2 \leq 1$

(C,N) if $y_1 \leq F_2(1)$ but $y_2 > 1$

(N,N) otherwise.

With one opportunity each (either simultaneously or sequentially) for a binary message, the following is an equilibrium:

(C,C) if both the y_i are ≤ 1

(N,N) otherwise.

[3] I am not wholly convinced of the legitimacy of equilibrium analysis. However, I use it for ease of comparison with other work such as Myerson (1983). Two alternative justifications for treating equilibria are: (i) perhaps these are agents drawn from a large population which has been engaging in this interaction for a long time - the "biological" approach - and (ii) perhaps, before coming to know their private information, the agents have non-bindingly agreed on how they will all behave. (Whether such an agreement is plausible is a question I address in related work.)

[4] Thus two players can converse secretly without others knowing the fact; or they can take care to be seen leaving the room together, or they can recapitulate some part of their discussion later in public,...

[5] Consider player I, and suppose he attaches probability p to the event

"II is S." Suppose first that I is S. If he announces "S," he will get no further information about II's type, and he expects II to play X. So the expected payoff from X is p, and that from Y is (1-p). Thus, having truthfully announced "S," I would play X as intended, and would get an expected payoff p ($p > 1-p$).

On the other hand, if I is in fact S but announces "W," there is probability p that he will find out that II has announced "S," which in equilibrium I will take to indicate that II is in fact S. Since I knows that II has no information about him (I), he will expect II to play X, and will therefore play X himself, getting payoff 1 in this case. There is a probability (1-p) that I will find out that II is W. In this case, II also believes that I is W, and so II will play Y, thus giving I payoff -2.

Since p is greater than $p + (1-p)(-2)$, when I is S he will report the truth. Now suppose that I is in fact W. If he reports W, then, with probability p, he discovers that II is S and that II will play X. This enables I to get 0 by playing X as he is meant to, and he would do no better by playing Y. With probability (1-p), he discovers that II is also W, and that II is going to play Y, as instructed; this makes it in I's interest also to follow the instruction to play Y, and he gets 1.

Suppose that when I is W he reports S. Then he will discover nothing about II's type, but will know II will play X. This gives I a payoff of 0 whatever II's type turns out to be. This is less than he can get by reporting honestly, i.e. 0 or 1 depending on II's type.

7. See Farrell (1982).

8. This is *not* an innocuous assumption. It is equivalent to assuming that communication is ruled out once the first move in G is made. I consider removing this restriction below.

9. I think this is no real loss of generality, since any randomizing device could be included in T_1, and also since I don't think 1 would ever choose to randomize, as it would be unobservable.

10. We could use the condition "weakly unconcealed" obtained by weakening the inequality in defining "unconcealed." But our purpose is to discover whether there are facts that will emerge even if the intention is for everyone to remain silent. Thus, we require strict inequality, representing an unwillingness to do what was intended.

11. Requiring this just for *some* $t_1 \in S_1$, with equality for the rest, would be inappropriate. If strict inequality holds for $t_1 \in S_1' \subseteq S_1$, with equality for $t_1 \in S_1 \setminus S_1'$, it is S_1' that would be inferred if S_1 were claimed. So we should consider whether the subset S_1' is credible and unconcealed.

12. For a related approach, based on exploitation of opportunities to "invent signals," see McLennan (1983).

13. The Difference vanishes if (as in Crawford-Sobel, for instance) each t is negligible in each coalition.

REFERENCES

Akerlof, G. (1970), "The Market for Lemmons: Qualitative Uncertainty and the Market Mechanism," Quarterly Journal of Economics, 84, 488-500.

Bernheim, D. (1982), "Rational Strategic Behavior," mimeo Stanford University.

Crawford, V., and J. Sobel (1982), "Strategic Information Transmission," Econometrica, 50, November.

Farrell, J. (1982), "Communication In Games," mimeo, MIT.

Goffman, E. (1969) Strategic Interaction, University of Pennsylvania Press.

Jervis, R. (1970), The Logic of Images In International Relations, Princeton University Press.

Kalai, E., and D. Samet (1983), "Persistent Equilibria in Strategic Games," CMSEMS discussion paper 515, Northwestern University.

Kreps, D., R. Wilson, P. Milgrom, and J. Roberts (1982), "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," Journal of Economic Theory.

Laffont, J.J., and E. Maskin (1982), "The Theory of Incentives: An Overview," in W. Hildenbrand, ed., "Advances in Economic Theory," Cambridge University Press.

McLennan, A. (1983), "Games with Communication," unpublished, University of Toronto.

Maynard Smith, J. (1974) "The Theory of Games and the Evolution of Animal Conflict," Journal of Theoretical Biology, 47.

Milgrom, P., and J. Roberts (1982), "Limit Pricing and Entry Under Incomplete Information: An Equilibrium Analysis," Econometrica, March.

Myerson, R. (1983), "Incentive Compatibility and Bayesian Equilibrium: An Introduction," CSMES discussion paper, Northwestern University.

Myerson, R. (1983b), "Mechanism Design by an Informed Principal," Econometrica, 51 (November).

Pearce, D. (1982), "Ex-ante Equilibria," mimeo, Princeton.

Spence, A.M. (1970), Market Signaling, Harvard University Press.

Wilson, C. (1980), "The Nature of Equilibrium in Markets with Adverse Selection," Bell Journal, Spring.