DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

EQUILIBRIUM SELECTION IN SIGNALING GAMES

Jeffrey S. Banks
California Institute of Technology

Joel Sobel
University of California, San Diego

SOCIAL SCIENCE WORKING PAPER 565

March 1985
Revised October 1985

ABSTRACT

We present a refinement of the set of sequential equilibria (Kreps & Wilson (1982)) for generic signaling games based on rationality postulates for off-the-equilibrium-path beliefs. This refinement concept eliminates equilibria which Kreps (1985) and others dismiss on intuitive grounds. In addition, we derive a characterization of the set of stable equilibria (Kohlberg and Mertens (1982)) for generic signaling games in terms of equilibrium strategies and restrictions on beliefs. Examples are given which differentiate the predictions of these equilibrium concepts.

EQUILIBRIUM SELECTION IN SIGNALING GAMES*

by Jeffrey S. Banks and Joel Sobel

I. INTRODUCTION

This paper investigates the relationship between Kreps and Wilson's (1982) concept of sequential equilibria and Kohlberg and Mertens's (1984) concept of stability. It introduces a restriction on off-the-equilibrium-path beliefs that refines the set of sequential equilibria in signaling games. We call any sequential equilibria that satisfies our restriction on beliefs divine. For generic signaling games, every equilibrium contained in a stable component is divine. Moreover, the solution concept is restrictive enough to rule out all of the equilibria that Kreps (1985) and others dismiss on intuitive grounds. Thus, divinity provides an independent theoretical foundation for discarding non-intuitive equilibria in signaling games.

We provide a generic example to show that divine equilibria may not be contained in any stable component. However, the paper presents an explicit characterization of stability in terms of off-the-equilibrium-path beliefs. That is, an equilibrium of a generic signaling game is in a stable component if and only if it can be supported as a sequential equilibrium with restricted off-the-equilibrium-path beliefs. Just as Kreps and Wilson (1982) characterize perfect equilibria for generic extensive-form games in terms of sequential equilibrium strategies and beliefs, our result characterizes stable outcomes for generic signaling games in terms of sequential equilibrium strategies and restrictions on beliefs. The characterization may be a useful way to compute stable equilibrium outcomes and to evaluate the consequences of using stability to select equilibria in extensive-form games.

Independent of our work, Cho and Kreps (1985) analyze the power of stability to select equilibria in signaling games. Their results closely parallel our own. They identify restrictions on equilibria similar to those embodied by divinity. In addition, they also state our characterization result (Theorem 3). Cho (1985) extends a restriction identified in Cho and Kreps to obtain a solution concept that refines the set of sequential equilibria in general extensive-form games.

Our debt to the existing literature on solution concepts for noncooperative games is obvious. Recent work on this topic includes papers by Kreps and Wilson (1982), Selten (1975), and McLennan (1985), who present refinement concepts for extensive-form games; and Myerson (1978), Kalai and Samet (1984), and Kohlberg and Mertens (1984), who present refinement concepts for normal-form games.

II. THE MODEL

In this paper we analyze the equilibria of signaling games with finite action sets. There are two players, a Sender (S) and a Receiver (R). The Sender has private information, summarized by his type, t, an element of a finite set T. There is a strictly positive

probability distribution p(t) on T; p(t), which is common knowledge, is the ex ante probability that S's type is t. After S learns his type he sends a message m to R; m is an element of a finite set M. In response to m, R selects an action a from a finite set A(m); k(m) is the cardinality of A(m). S and R have von Neumann-Morgenstern utility functions u(t,m,a) and v(t,m,a), respectively.

For fixed T, M, and A(m) for m ε M, the utility functions u(t,m,a) and v(t,m,a) completely determine the game. Therefore, if $L = [\bar{T} \times \sum_{i=1}^{\bar{M}} k(i)]^2$, where $\bar{T}$ is the cardinality of T and $\bar{M}$ is the cardinality of M, then every element of $\mathbb{R}^L$ determines a signaling game. We call a property of a signaling game generic if there exists $D \subset \mathbb{R}^L$ such that the property holds for all signaling games determined by d ε D and a closed set of Lebesgue measure zero contains $\mathbb{R}^L \backslash D$. If a property of a signaling game is generic, then we say it holds for generic signaling games.

For any positive integer k, let $\Delta_k = \{\delta = (\delta(1),\ldots,\delta(k)): \delta(i) \geq 0 \quad \forall\, i \text{ and } \sum_{i=1}^{k} \delta(i) = 1\}$ be the (k - 1)-dimensional simplex. We refer to the $(\bar{T} - 1)$-dimensional simplex most often; to simplify notation, we write Δ instead of $\Delta_{\bar{T}}$. A signaling rule for S is a function

$$q: T \to \Delta_{\bar{M}};$$

q(m|t) is the probability that S sends the message m, given that his type is t. An action rule for R is an element of $\prod_{m \in M} \Delta_{k(m)}$;

r(a|m) is the probability that R uses the pure strategy a when he receives the message m.

We extend the utility functions u and v to the strategy spaces $\Delta_{k(m)}$ by taking expected values; for all t ε T, let

$$u(t,m,r(\cdot)) = \sum_{a \in A(m)} u(t,m,a)\, r(a|m)$$

$$v(t,m,r(\cdot)) = \sum_{a \in A(m)} v(t,m,a)\, r(a|m).$$

Also, for each λ ε Δ and m ε M let

$$BR(\lambda,m) \equiv \arg\max_{r(m) \in \Delta_{k(m)}} \sum_{t \in T} v(t,m,r(m))\lambda(t)$$

be the best-response correspondence for R.

Definition. A sequential equilibrium for a signaling game consists of signaling rules q(t) for S, action rules r(m) for R, and beliefs μ(·|m) ε Δ for R, such that

1) $\forall\, t \in T$, $q(m^*|t) > 0$ only if

$$u(t,m^*,r(m^*)) = \max_{m \in M} u(t,m,r(m));$$

2) $\forall\, m \in M$, $r(a^*|m) > 0$ only if

$$\sum_{t \in T} v(t,m,a^*)\mu(t|m) = \max_{a \in A(m)} \sum_{t \in T} v(t,m,a)\mu(t|m);$$

3) if $\sum_{t \in T} q(m|t)p(t) > 0$, then

$$\mu(t^*|m) = \frac{q(m|t^*)p(t^*)}{\sum_{t \in T} q(m|t)p(t)} .$$

In words, (1) states that q(·) maximizes S's expected utility, given R's strategy; (2) states that r(·) maximizes R's expected utility, given beliefs μ(·); and (3) states that R's beliefs given S's strategy are rational in the sense that Bayes' Rule determines μ(t|m) whenever the probability that S sends m in equilibrium is positive. If q(m|t) = 0, for all t ε T, then sequential rationality does not determine μ(t|m). However, the refinement concept introduced in Section 3 restricts the values that these beliefs may take.

Next, we describe stable equilibria. Our introduction follows Kreps (1985). Fix a signaling game; let $\tilde{\rho} = (\tilde{\rho}_R, \tilde{\rho}_S)$ satisfy $0 < \tilde{\rho}_i < 1$, i = R,S, and let $\tilde{q}$ and $\tilde{r}$ be strategies for S and R respectively that satisfy $\tilde{q}(m|t) > 0$, ∀ m ε M, ∀ t ε T and $\tilde{r}(a|m) > 0$, ∀ a ε A(m), ∀ m ε M. A $(\tilde{\rho},\tilde{q},\tilde{r})$-perturbation of the original game is the signaling game in which, if the players choose strategies q and r from the original game, then the outcome is the outcome of the original game if the strategy chosen by S is $(1 - \tilde{\rho}_S)q + \tilde{\rho}_S\tilde{q}$ and the strategy chosen by R is $(1 - \tilde{\rho}_R)r + \tilde{\rho}_R\tilde{r}$. We refer to $(\tilde{\rho},\tilde{q},\tilde{r})$ as trembles. Let (q,r) be Nash equilibrium strategies for a perturbed game. If q(m|t) > 0, we say that a type t Sender voluntarily sends m and we say that R voluntarily uses the mixed strategy r(m).

For a given signaling game, we call a subset C of the set of Nash equilibria stable if, for every ε > 0 there exists δ > 0 such that every $(\tilde{\rho},\tilde{q},\tilde{r})$-perturbation of the original game with $0 < \tilde{\rho}_i < \delta$, i = R,S has an equilibrium no more than ε from the set C.

Definition. A stable component is a minimal (by set inclusion) stable set of equilibria.

Our analysis depends on several properties.[1]

Proposition 1. For generic extensive-form games, the set of equilibrium probability distributions on endpoints[2] is finite and all equilibria within a given connected component induce the same probability distributions on endpoints.

Proposition 2. Every game has at least one stable component.

Proposition 3. A stable set of equilibria remains so when one deletes a strategy that is not a best reply against any equilibrium in the set.

Therefore, in generic signaling games, there exists a stable set of equilibria with the property that every equilibrium in the set agrees along the equilibrium path; the equilibrium may vary off the equilibrium path. A variety of off-the-equilibrium-path responses may be needed to guarantee that any perturbation of the game has an equilibrium path close to a particular equilibrium path. Therefore, a

single equilibrium need not be a stable set. However, we use Proposition 1 to justify an abuse of terminology. We call an equilibrium stable if it agrees with an element of a stable component along the equilibrium path. In particular, in generic signaling games, if an equilibrium is stable, then every perturbation has an equilibrium with payoffs close to the original equilibrium payoffs.

III. DIVINE EQUILIBRIA

Previous refinements of the Nash equilibrium concept place rationality restrictions on zero-probability events. In particular, sequential rationality requires that players respond optimally to some consistent assessment of how the game has been played. These equilibrium concepts do not require a player to draw any conclusion when a zero-probability event takes place. That is, although the refinement concepts embodied in sequential rationality and perfectness require that equilibria of games induce equilibria on any continuation of the game, these concepts do not require that a player systematically draw an inference from an opponent's unexpected move. Nevertheless, in order to decide how to respond to an unexpected signal, R should evaluate the willingness of S-types to deviate from equilibrium, and then incorporate into his beliefs the information that deviations from equilibrium might reveal.

This section presents an equilibrium concept that refines the set of sequential equilibria in signaling games by placing restrictions on off-the-equilibrium-path beliefs. We begin by describing two restrictions on beliefs along with the intuition behind them, and then proceed to define an equilibrium concept that incorporates these restrictions.

The first intuitive restriction on beliefs that we discuss requires R's off-the-equilibrium-path beliefs to place positive probability only on those Sender types who might not lose from a defection. Formally, this condition requires that if

$J = \{t: u^*(t) > u(t,m,r(m)) \text{ for all } r(m) \in BR(\Delta,m)\}$,

then $r^*(m) \in BR(\Delta_{T\setminus J},m)$.[3] Cho and Kreps (1985) also identify this condition and show that if an equilibrium is stable, then the condition must hold.[4]. Our refinement notion includes this type of restriction on beliefs.

Figure 1[5] describes a special case of a sequential settlement game (see Salant (1984) or Sobel (1985)). There are two types of S (the "defendant"): type t_2 defendants are negligent; type t_1 defendants are not negligent. S offers a low settlement, m_1, or a high settlement, m_2. R (the "plaintiff") either accepts (a_1) or rejects (a_2) the offer. If R accepts S's offer, S pays R an amount that depends only on the offer. If R rejects the offer, S must pay court costs and a transfer depending only on his type (e.g. the court finds out with certainty whether or not S was negligent). If $p(t_1) = p(t_2) = \frac{1}{2}$, then the game depicted in Figure 1 has two types of equilibria.

m_1	a_1	a_2
t_1	-3,3	-6,0
t_2	-3,3	-11,5

m_2	a_1	a_2
t_1	-5,5	-6,0
t_2	-5,5	-11,5

Figure 1

In one type of equilibrium, both types of S offer m_1, and R accepts any offer; $q(m_1|t_i) = 1$, $i = 1,2$, $r(a_1|m_j) = 1$, $j = 1,2$. In the other type of equilibrium, both types of S offer m_2 and R accepts m_2 and rejects m_1; $q(m_1|t_i) = 0$, $i = 1,2$, $r(a_1|m_1) = 0$, $r(a_1|m_2) = 1$. In order to support this behavior, we need $\mu(t_1|m_1) \leq \frac{2}{5}$. We claim that the second equilibrium is not plausible because, in order to support it, R must believe that t_2 is more likely than t_1 to offer m_1. However, t_1 prefers to defect whenever t_2 does (and not conversely: consider an equal mixture of a_1 and a_2 given m_1). Thus, a reasonable restriction on beliefs would require that the relative probability of t_1 should increase if R observes m_1. Our refinement notion captures this argument as well.

Fix an equilibrium in which a Sender of type t obtains utility $u^*(t)$, and, for all $t \in T$, the probability that t sends m is zero. We intend to restrict the beliefs that R can have given the message m. Since we deal with only one unsent message at a time, for notational convenience we drop the argument m from R's response function.

Recall that $\Delta_{k(m)}$ consists of all actions, r, available to R given m. Let

$$A_G = \{r \in \Delta_{k(m)}: u(t,m,r) \geq u^*(t), \text{ for some } t \in T\}$$

be the set of actions that some S-type weakly prefers to equilibrium actions, conditional on sending m. Our initial restriction is that R should believe that any type who sends m instead of the equilibrium signal does not expect to lose by doing so.[6] Thus, if R receives the signal m (as a defection from equilibrium), he should believe that S expects him to take an action in A_G.

For all $r \in \Delta_{k(m)}$, let

$$\bar{\mu}(t,r) = \begin{cases} 1 & \text{if } u(t,m,r) > u^*(t) \\ [0,1] & \text{if } u(t,m,r) = u^*(t) \\ 0 & \text{if } u(t,m,r) < u^*(t) \end{cases}$$

be the frequency that $t \in T$ would send m if he believes that m would induce the action r and t had a choice between sending m or obtaining $u^*(t)$. Next, let

$$\Gamma(r) = \{\gamma \in \Delta: \exists\ \mu(t) \in \bar{\mu}(t,r) \text{ and } c > 0 \text{ such that } \gamma(t) = c\mu(t)p(t), \ \forall\ t \in T\}.$$

Notice that $\Gamma(r)$ is nonempty if and only if $r \in A_G$. If it is common knowledge that m induces r, then the posterior probability distribution over T must be an element of $\Gamma(r)$. Thus, $\Gamma(r)$ is the set of beliefs consistent with R taking the action r in response to m (and t earning $u^*(t)$ otherwise).

Finally, let

$$\bar{\Gamma}(A) = \text{convex hull}\left[\bigcup_{r \in A} \Gamma(r)\right].$$

Thus, if A is closed, then $\bar{\Gamma}(A)$ is a closed, convex subset of the simplex Δ, and is empty if and only if $A_G \cap A$ is empty. Since $\bar{\Gamma}(\Delta_{k(m)})$ is empty only if $u^*(t) > u(t,m,r)$, $\forall\ t \in T$, $\forall\ r \in \Delta_{k(m)}$, R truly would be surprised by a defection from equilibrium, and there seems to be no reason to select one inference over another in response to m. Indeed, in this case, <u>any</u> conjecture supports the equilibrium. When $A_G \neq \emptyset$, and hence $\bar{\Gamma}(\Delta_{k(m)}) \neq \emptyset$, we think that it is not plausible for R to hold beliefs outside of $\bar{\Gamma}(\Delta_{k(m)})$ given the signal m. If R observes a defection from the equilibrium path, then he must form a conjecture over T <u>based on that defection</u>.

Notice that any equilibrium in which beliefs lie in $\bar{\Gamma}(\Delta_{k(m)})$ satisfies the intuitive restrictions that we described earlier. All conjectures in $\bar{\Gamma}(\Delta_{k(m)})$ assign zero probability to any $t \in T$ with $u(t,m,r) < u^*(t)$, $\forall\ r \in \Delta_{k(m)}$. Furthermore, if there exists $t,t' \in T$ such that $\bar{\mu}(t,r) = 1$ implies $\bar{\mu}(t',r) = 1$, $\forall\ r \in \Delta_{k(m)}$, then for all beliefs in $\bar{\Gamma}(\Delta_{k(m)})$, the ratio of the probability of t' given m to the probability of t given m is at least as great as $\frac{p(t')}{p(t)}$. That is, R believes that t' is at least as likely to defect as t.

Beliefs must lie in $\bar{\Gamma}(\Delta_{k(m)})$ provided two conditions hold. First, R believes that no type t would use m if t expected R to take an action that resulted in utility less than $u^*(t)$. This means that S expects R to take actions in A_G given the signal m. Second, S-types have a common conjecture over the distribution of actions that R would take as a response to a defection. This second condition may seem odd, since there is only one Sender. However, a "type" is a specification

of the information S has concerning decision parameters that are not common knowledge. Thus, it is possible for two S-types to have different conjectures over R's actions in equilibrium. If it is common knowledge that R holds beliefs in $\bar{\Gamma}(\Delta_{k(m)})$, then S should expect m to induce an action in $BR(\bar{\Gamma}(\Delta_{k(m)}),m)$. This observation suggests the following iterative procedure. Let

$$\Gamma'_0 = \Delta, \quad A_0 = \Delta_{k(m)}, \quad \text{and for } n > 0,$$

$$\Gamma'_n = \begin{cases} \bar{\Gamma}(A_{n-1}) & \text{if } \bar{\Gamma}(A_{n-1}) \neq \emptyset \\ \Gamma'_{n-1} & \text{if } \bar{\Gamma}(A_{n-1}) = \emptyset \end{cases}$$

$$A_n = BR(\Gamma'_n,m), \quad \Gamma^* = \bigcap_n \Gamma'_n, \quad A^* = \bigcap_n A_n .$$

Others use iterative procedures in the definition of equilibrium concepts. Specifically, given the assumptions that S expects R to take actions in A_G given an unexpected signal m and that S-types have a common conjecture over the actions that R would take in response to m, our iterative procedure coincides with that used by Bernheim (1984) and Pearce (1984) to define the set of rationalizable equilibria.

<u>Theorem 1</u>. In generic signaling games, if an equilibrium in which $q(m|t) = 0$ $\forall\ t \in T$ is stable, then there exists $r^* \in A^*$ such that $u(t,m,r^*) \leq u^*(t)$, $\forall\ t \in T$.

Theorem 1 is a direct consequence of Proposition 3. It states that if an equilibrium is stable, then there exist beliefs in Γ^* that support it. We discuss the proof later in this section.

Definition. A sequential equilibrium in a signaling game is divine if it is supported by beliefs in Γ^*.

Thus, by Theorem 1, every stable component contains a divine equilibrium. Therefore, Proposition 2 implies our next result.[7]

Theorem 2. Every signaling game has a divine equilibrium.

We believe that divinity captures a minimal restriction on off-the-equilibrium path beliefs. Stability implies much more, but we are not convinced that these restrictions are plausible.

The set of beliefs in Γ^* depend on the prior distribution of Sender types. To check this property, one need only note that in the game that Figure 1 describes,

$$\Gamma^* = \{\lambda \in \Lambda: \lambda(t_1) \geq p(t_1)\}$$

for the equilibrium in which both t_1 and t_2 send m_2 with probability one. Let Γ^{**} be the intersection of the Γ^* taken over all nondegenerate priors on Sender types. We can show that in generic signaling games, if an equilibrium is stable, then it can be supported by beliefs in Γ^{**}. Call an equilibrium supported by beliefs in Γ^{**} universally divine. To see that universal divinity is more restrictive than divinity alone, note that in Figure 1, the sequential

equilibrium in which S sends m_2 with probability one is divine provided that $p(t_1) \leq \frac{2}{5}$, but it is never universally divine since, regardless of the prior probability that S is t_1, R must believe that the unexpected signal m_1 comes from t_1.[8]

Cho and Kreps use Proposition 3 to further refine the equilibrium set. For a fixed equilibrium outcome and unsent signal m, call a type t bad for m if, for every equilibrium giving rise to this outcome, a t-Sender strictly prefers the equilibrium outcome to sending m.[9] Proposition 3 implies that a stable equilibrium can be supported by beliefs that give no weight to any type that is bad for m (if all types are bad for m, then the equilibrum payoffs strictly dominate any payoff S can obtain from a best response to m). To see that this condition is more restrictive than universal divinity, note that for generic signaling games, if t is not bad for m, then e(t), the element of Δ with t-th component equal to one, is an element of Γ^{**}.[10] Thus, Proposition 3 also implies that in generic signaling games, if an equilibrium is stable, then there exist beliefs in Γ^{**} that support it. Since $\Gamma^{**} \subset \Gamma^*$, Theorem 1 follows from Proposition 3.

IV. A CHARACTERIZATION OF STABLE EQUILIBRIA

This section gives necessary and sufficient conditions for a sequential equilibrium in a generic signaling game to be stable. First, we present an example of a signaling game that has an unstable, divine equilibrium. The example motivates the notion of stable beliefs that we need to prove our equivalence theorem.

Consider the signaling game in Figure 3.

m_1	a
t_1	0,0
t_2	0,0

m_2	a_1	a_2	a_3	a_4
t_1	-1,3	-1,2	1,0	-1,-2
t_2	-1,-2	1,0	1,2	-2,3

Figure 3

Let $p(t_1) = \frac{1}{2}$. There exists a sequential equilibrium to this game in which $q(m_1|t_1) = 1$, $i = 1,2$, $r(a_1|m_2) = 1$ supported by beliefs $\mu(t_1|m_2) \geq \frac{2}{3}$. This equilibrium is universally divine since

$\Gamma^* = \Gamma^{**} = \Delta$ and

$a_1 \in BR(\Gamma^*, m_2)$; also, neither t_1 nor t_2 is bad for m_2 so that the Proposition 3 does not restrict beliefs. However, this equilibrium is not stable.

The stable equilibrium for this example involves both t_1 and t_2 sending m_2 with probability one and R responding to m_2 with actions a_2 and a_3 with probability $\frac{1}{2}$ each.

Now we argue that the equilibrium in which S does not use m_2 is not stable. Notice that if S voluntarily sends m_2 an equilibrium to the perturbed game in which S types expect to receive 0, then R must either use an equal mixture of a_1 and a_2 or an equal mixture of a_3 and a_4 in response to m_2. Hence, R must believe that the probability of t_1 given m_2 is equal to either $\frac{2}{3}$ or $\frac{1}{3}$. Any other

strategy for R leads to positive payoffs for at least one S type or negative payoffs to both. Moreover, when R mixes equally between a_1 and a_2, t_1 does not voluntarily send m_2 and when R mixes equally between a_3 and a_4, t_2 does not voluntarily send m_2. This argument establishes that if $\mu(t_1|m_2)$, the probability of t_1 given m_2 if S does not voluntarily send m_2, is an element of $(\frac{1}{3},\frac{2}{3})$, then there is an equilibrium to the perturbed game close to the original equilibrium only if the tremble induces R to take an action given m_2 that does not attract either type of S. Moreover, if $\mu(t_1|m_2) \notin (\frac{1}{3},\frac{2}{3})$, then the perturbed game has an equilibrium that is close to the original game and in which either t_1 or t_2 voluntarily sends m_2. Therefore, the equilibrium in the example is stable if and only if, given m_2, every best response to the set of beliefs in which the probability of t_1 given m_2 is an element of $(\frac{1}{3},\frac{2}{3})$ leads to nonpositive expected payoffs to both S types. Since $a_3 \in BR((\frac{1}{3},\frac{2}{3}),m_2)$ yields positive payoffs to both S types, the equilibrium is not stable. We apply an analogous argument in general signaling games. First, we identify the set of trembles that cannot induce voluntary action in any equilibrium to the perturbed game that is close to the original equilibrium. Second, we prove that an equilibrium is stable precisely when no best response to this set of trembles induces S to voluntarily send m.

As in the previous section, fix an equilibrium that leads to utility levels $u^*(t)$, $\forall\, t \in T$, and in which $q(m|t) = 0$, $\forall\, t \in T$. For each $J \subset T$, define

$$I(J) \equiv \{r \in \Delta_{k(m)}: u^*(t) \geq u(t,m,r) \ \forall\, t \in T, \text{ and}$$
$$u^*(t) = u(t,m,r) \text{ if and only if } t \in J\},$$

and, for $r \in I(J)$, define

$$\hat{\Lambda}(J,r) \equiv \{\lambda \in \text{int } \Delta: \nexists\, \lambda^* \in \Delta \text{ with } r \in BR(\lambda^*,m)$$

$$\text{such that } \lambda^* = \sum_{t \in J} \alpha(t)e(t) + \beta\lambda, \text{ for}$$

$$\alpha(t) \geq 0, \quad 1 - \sum_{t \in J} \alpha(t) \equiv \beta > 0\},$$

where $e(t) \in \Delta$ is the vector with t-th component equal to one and all other components equal to zero. Finally, let

$$\Lambda(J) \equiv \begin{cases} \bigcap_{r \in I(J)} \hat{\Lambda}(J,r) & \text{if } I(J) \neq \emptyset \\ \Delta & \text{if } I(J) = \emptyset \end{cases}$$

and $\Lambda^* = \bigcap_{J \subset T} \Lambda(J)$.

Consider a perturbed game in which trembles induce a belief λ given m unless some type voluntarily uses m. For sufficiently small trembles, there exists an equilibrium to the perturbed game, with payoffs close to $u^*(t)$, in which R takes action r given m if and only if $\lambda \notin \hat{\Lambda}(J,r)$ for some J; the action r is not a best response to any beliefs obtained by "adding" combinations of $t \in J$ to λ if and only if $\lambda \in \hat{\Lambda}(J,r)$. As only S-types in J voluntarily use m in an equilibrium in which they could obtain $u^*(t)$ by not sending m, $\hat{\Lambda}(J,r)$ contains

exactly the beliefs that may cause instability if R takes action r given m. Thus, $\bigcap_{J \neq \emptyset \subset T} \Lambda(J)$ is the set of trembles that cannot induce voluntary action in any equilibrium. However, $\Lambda(\emptyset)$ are those beliefs which give rise to actions attractive to some S types. This argument leads to our characterization theorem.

Theorem 3. In generic signaling games, an equilibrium is stable if and only if, for all unused signals m, $\Lambda^* = \emptyset$.

V. EXTENSIONS

While we confine our discussion in this paper to signaling games, Propositions 1-3 hold for generic extensive-form games. Since these results combine to imply Theorems 1 and 2, we can use our techniques to rule out implausible sequential equilibria in more general extensive-form games. We suspect that divinity is easier to verify than stability and may be simpler to generalize to games with infinite strategy spaces. On the other hand, Theorem 3 and possible generalizations appear to be valuable only as a characterization of stable equilibria.

We conclude by noting that our techniques do not refine the set of sequential equilibria in signaling games in which signals are costless. Specifically, let A(m), u(t,m,a), and v(t,m,a) be independent of m. These games are not generic, so we cannot apply our results directly. However, it is easy to verify that $\Gamma^* = \Delta$ for any unused signal. This is because if t induces the action $a \in \bar{A}$ with signal m', then there exist beliefs for which a is a best response to

the (unused) signal m. When signaling is costless, t is indifferent between sending m and m′ and no other agent strictly prefers m to his equilibrium payoff. In addition, straightforward arguments show that stability does not restrict the set of equilibria, although this kind of game always has an equilibrium in which all types of S send the same signal and typically has other, more appealing, equilibria. Farrell[11] (1984) and Myerson (1983) present ideas that apply to costless signaling games. Myerson presents an axiomatic solution that limits the outcomes in a mechanism-design problem that usually has a large number of sequential equilibria, but it is not clear that his ideas extend in a sensible way to a noncooperative framework. Farrell argues that an equilibrium outcome is not plausible if there exists an unused signal m, a nonempty set J, and an action $r \in BR(\lambda,m)$ such that

$J = \{t: u^*(t) < u(t,m,r)\}$, where

$$\lambda(t) = \begin{cases} p(t)/\sum_{t' \in J} p(t') & \text{if } t \in J \\ 0 & \text{if } t \notin J \end{cases}$$

is the conditional probability of t given $t \in J$. That is, Farrell argues that R should interpret a defection that benefits exactly the set J as evidence that exactly those t in J use m. Farrell calls an equilibrium in which this type of defection does not exist neologism proof. Neologism-proof equilibria do not exist in general, and, in games with costly signaling, need not be divine.

NOTES

* Presented at the 5th World Congress of the Econometric Society, Boston MA, August 1985. We thank participants of Caltech, UCSD and Rand Corporation Theory Workshops, Drew Fudenberg, David Kreps, and two referees for valuable comments. Sobel thanks Joe Farrell and Chris Harris for many conversations on related topics and the National Science Foundation for partial support under grant SES 84-08655.

1. Kreps and Wilson (1982) prove Proposition 1. Kohlberg and Mertens (1984) prove Porpositions 1-3.

2. An equilibrium induces a probability distribution on the endpoints of the tree. An equilibrium probability distribution on endpoints is a probability distribution on endpoints induced by some equilibrium.

3. If J = T, then no action R can take in response to the signal m induces S to send m. In this case, any beliefs are permissable.

4. Kreps (1985) suggests a less restrictive version of this condition. Kreps discards an equilibrium in which there exists a sender type who would like to defect for every action in $BR(\Delta_{T\setminus J},m)$.

5. We represent examples with a bi-matrix B(m) for each $m \in M$. There is one column in B(m) for each strategy in A(m) and one row for

each type. The entry in the t-th row and the a-th column is (u(t,m,a), v(t,m,a)). In each of these examples, the qualitative properties that we discuss in the text remain valid if we perturb the entries in B(m).

6. It does not change our results to require that R believes that any type who sends m instead of the equilibrium signal expects to benefit strictly by doing so. Thus, we can use a strong inequality in the definition of A_G.

7. Strictly speaking, Theorem 1 and Proposition 3 imply the existence of divine equilibria in generic signaling games. A limiting argument, based on the upper hemi-continuity of divine equilibrium paths, establishes Theorem 2. Cho (1985) gives the details of a related argument.

8. Harris and Raviv (1983) study a game in which there is a divine equilibrium that is not universally divine, hence not stable. Their comparative-statics analysis concentrates on the stable path.

9. McLennan (1985) defines a refinement concept that is similar in spirit to this requirement. Specifically, call an action <u>useless</u> if it has a suboptimal payoff in every sequential equilibrium of a game (not just those equilibria in a stable component). McLennan shows that there exist sequential equilibria with beliefs restricted so that, at each information set, they assign positive

probability only to nodes reached by the fewest useless actions. From this, McLennan recursively defines higher-order uselessness and arrives at a set of <u>justifiable</u> equilibria. In generic signaling games, only strongly dominated actions are useless, thus any divine equilibrium is justifiable.

10. This condition is strictly more restrictive than universal divinity. In the game described in Figure 2, there is a sequential equilibrium in which both S types send m_1 with probability one and R takes a_3 given m_2. It is straightforward to check that $\Gamma^{**} = \Delta$. However, the message m_2 is bad for t_2. When R believes only t_1 would send m_2, R's best response given m_2 is a_1. Therefore, the equilibrium is not stable.

m_1	a
t_1	0,0
t_2	0,0

m_2	a_1	a_2	a_3	a_4
t_1	-1,3	1,2	-1,0	1,-2
t_2	1,-2	1,0	-2,2	-1,3

Figure 2

11. Grossman and Perry's (1984) concept of perfect sequential equilibria is similar to Farrell's concept. However, Grossman and Perry analyze a particular game with costly signaling.

REFERENCES

Bernheim, D. "Rationalizable Strategic Behavior." Econometrica 52 (1984):1007-28.

Cho, I. "A Refinement of the Sequential Equilibrium Concept." Stanford mimeo, 1985.

Cho, I. and Kreps, D. "More Signalling Games and Stable Equilibria." Stanford mimeo, 1985.

Farrell, J. "Credible Neologisms in Games of Communication." MIT mimeo, 1984.

Grossman, S. and Perry, M. "Sequential Bargaining under Asymmetric Information." Foerder Institute Working Paper 33-84, 1984.

Harris, M. and Raviv, A. "A Sequential Signalling Model of Convertible Debt Call Policy." IMSSS Working Paper, 1983.

Kalai, E. and Samet, D. "Persistent Equilibria." International Journal of Game Theory 13 (1984):129-44.

Kohlberg, E. and Mertens, J.-F. "On the Strategic Stability of Equilibria." Harvard Business School Working Paper 1-785-012, 1984.

Kreps, D. "Signalling Games and Stable Equilibria." Stanford mimeo, 1985.

Kreps, D. and Wilson, R. "Sequential Equilibria." Econometrica 50 (1982):863-94.

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

Myerson, R. "Refinement of the Nash Equilibrium Concept." International Journal of Game Theory 7 (1978):73-80.

_____. "Mechanism Design by an Informed Principal." Econometrica 51 (1983):1767-98.

Pearce, D. "Rationalizable Strategic Behavior and the Problem of Perfection." Econometrica 52 (1984):1029-50.

Salant, S. "Litigation of Settlements Demands Questioned by Bayesian Defendants." Caltech Social Science Working Paper 516, 1984.

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

Sobel, J. "An Analysis of Discovery Rules," UCSD Discussion Paper, 1985.