JOURNAL OF ECONOMIC THEORY 57, 36–51 (1992)

Signaling Future Actions and the Potential for Sacrifice*

ELCHANAN BEN-PORATH

MEDS Department, J. L. Kellog Graduate School of Management, Northwestern University, Evanston, Illinois 60208

AND

EDDIE DEKEL*

Department of Economics, University of California, Berkeley, California 94720

Received May 17, 1989; revised August 24, 1991

We consider extensions of games where some players have the option of signaling future actions by incurring costs. The main result is that in a class of games, if one player can incur costs, then forwards induction selects her most preferred outcome. Surprisingly, the player does not have to incur any costs to achieve this—the option alone suffices. However, when all players can incur costs, one player's attempt to signal a future action is vulnerable to a counter-signal by the opponent. This vulnerability to counter-signaling distinguishes signaling future actions from signaling types. *Journal of Economic Literature* Classification Numbers: 026.

1. INTRODUCTION

The idea of using costly signals to convey private information has been researched extensively at both the theoretical level and in applications. Spence [17] has shown that a worker might pay for education even when it has no real value in order to signal that she is competent. Cho and Kreps [5] studied a general class of such "signaling games," in which a player

* This is a revision of "Coordination and the Potential for Self Sacrifice," first draft dated November 1987. We thank an associate editor, a referee, and Matthew Rabin for detailed and helpful comments. Financial support from the Miller Institute, IBER, the Sloan Foundation and NSF Grant SES-8808133 are gratefully acknowledged. This work was begun while the first author was at the Graduate School of Business, Stanford University.

0022-0531/92 $5.00

	L	R
U	5,1	0,0
D	0,0	1,5

FIGURE 1.1

incurs costs to signal her private information about a move by nature, i.e., her *type* (Harsanyi [10]). Cho and Kreps showed that the idea of forwards induction is useful for selecting among the multiple equilibria which often exist in these signaling games. In this paper we apply forwards induction to show that the option of incurring a cost can signal a player's future actions. It turns out that, in contrast to signaling games, future actions can be signaled without any costs being incurred.[1]

The signaling of future actions can be demonstrated in a simple example. Consider the "battle of the sexes" game in Fig. 1.1. The outcome (U, L) is preferred by player 1 (the row player) to any other outcome, and is a strict Nash equilibrium. Suppose we extend the game to include a *signaling* stage, where player 1 has the possibility of burning, say, 2 units of utility before the game begins. This creates the extended game in Fig. 1.2. Burning and then playing D is strongly dominated for player 1 (by not burning and playing D) hence if player 2 observes 1 burning, then 2 can conclude that 1 will play U. Therefore player 1 can guarantee herself 3 by burning and

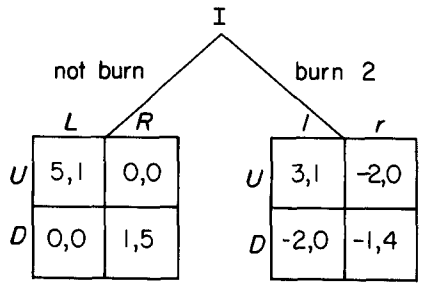

FIGURE 1.2

[1] Saying that future actions are "signaled" is not precise because no costs are actually incurred. In fact, the "receiver" deduces the "sender's" future action because the sender *could* have signaled. Nevertheless, we adopt the signaling terminology since it is more convenient than more precise alternatives, such as "the players deduce the future action of the player who has the potential to signal."

playing U, since 2 (having concluded that 1 will play U after burning) will play L. Now, we claim that even if player 1 does not burn, player 2 should conclude that 1 will play U. This is because, by playing D, player 1 can receive a payoff of at most 1, while the preceding argument demonstrated that player 1 can guarantee 3 (by burning). Hence, if 2 observes that 1 does not burn then 2 will play L, leading to player 1's preferred outcome which involves no burning and (U, L). The intuitive notion underlying this discussion is that of forwards induction, introduced by Kohlberg and Mertens [12]. The solution concept underlying the argument above is iterative deletion of weakly dominated strategies.

The main result of this paper is a generalization of the above example. In Section 2 we show that if one player can signal then she will attain her most preferred outcome whenever she has a unique best outcome that is a strict Nash equilibrium. Since the player who signals achieves her most preferred outcome, an important question is who can signal, and when. Occasionally the environment being modeled contains the answer to this question: e.g., if only one player can send costly signals.[2] But usually both players have the option of incurring costs. So the receiver can "counter-signal." We argue that if the receiver is given the option to counter-signal then it is no longer true that the sender can signal future actions. Thus signaling future actions is not robust to modifying the game by allowing for counter-signaling. On the other hand, signaling types is (trivially) robust to this modification. It is worth noting that *any* signaling of future actions, and not only signling by having the option of incurring costs, is vulnerable to counter-signaling. So, for example, signaling by forgoing an option (as in Kohlberg and Mertens [12, Section 2.3]) can also be invalidated by counter-signaling. We conclude that in many situations signaling future types is possible, while signaling future actions is, at the least, problematic.

Van Damme [18] also suggested the possibility of introducing into a game a stage where players may burn money, and provided an example similar to that described above.[3] The reader is urged to read his paper which presents many other interesting examples that help clarify the relationship between forwards induction and stable equilibrium. The role of forwards induction in the context of signaling games has been examined by Cho and Kreps [5], Banks and Sobel [2], Milgrom and Roberts [13], and others; and in other types of games by Cho [4], Glazer and Weiss

[2] For example, in the Spence [17] signaling model it may be that only the worker can signal. In the Appendix we show that a Spence signaling model can be modified so that the worker can use education to signal both her type and her action in a subsequent bargaining stage.

[3] We learned of van Damme's paper, which was written before ours, after the results and a draft of this paper had been completed. Our research was conducted independenty.

[9], and Osborne [15].[4] A different approach to selecting outcomes in games is developed in Farrell [7]. He introduces a language, i.e., *costless* communication together with explicit assuptions on how the "suggestions" of one player are interpreted by the others, and examines the implications on the outcome of the game when only one player can talk.[5]

2. Signaling Future Actions by Burning Money

We focus on two-person games and assume that one of the players, say 1, can burn utility before the play of a game G. We first formalize the extended game with signaling by burning. Let $G = \langle S, T, u, v \rangle$ where S, T are the strategy sets, and u, v are the utility functions, of players 1, 2, respectively. For simplicity we assume that player 1 can burn non-negative integer multiples of some positive number ε. The extended game is denoted by

$$G(\varepsilon) = \langle N \times S, T^N, \bar{u}, \bar{v} \rangle,$$

where N denotes the natural numbers, $\bar{u}((n, s), \mathbf{t}) \equiv u(s, \mathbf{t}(n)) - n\varepsilon$, and $\bar{v}((n, s), \mathbf{t}) \equiv v(s, \mathbf{t}(n))$ for any strategies $(n, s) \in N \times S$, $\mathbf{t} \in T^N$. It is convenient to think of $G(\varepsilon)$ in terms of the following description (see Fig. 2.1). Player 1 first chooses how much to burn ($n\varepsilon$); player 2 then observes the burning; and finally, the players simultaneously choose actions in the original game G (say (s, t)). The payoffs are then determined from G, where 1's payoff is decreased by the amount she burned: player 1 receives $u(s, t) - n\varepsilon$, and player 2 gets $v(s, t)$. For any set X, let $\Delta(X)$ denote the set of probability measures on X. Mixed strategies can be identified with behavioral strategies which are denoted by $(\eta, \boldsymbol{\sigma}) = (\eta, \boldsymbol{\sigma}(1), \boldsymbol{\sigma}(2), \ldots) \in \Delta(N) \times [\Delta(S)]^N$ for player 1 (where $\boldsymbol{\sigma}(n)$ is what 1 plays if she burns $n\varepsilon$ in the first stage), and $\tau = (\tau(1), \tau(2), \ldots) \in [\Delta(T)]^N$ for player 2. For notational simplicity $u: \Delta(S) \times \Delta(T) \to \mathfrak{R}$ denotes the extension of player 1's

[4] This paper is also related to other work on equilibrium selection. Aumann and Sorin [1] show that if the Pareto optimal outcome is unique then it is the only equilibrium in a repeated game where one player is uncertain about his opponents' type, and the possible types have bounded memories. Kalai and Samet [11] show that a refinement of Nash equilibrium yields the mutualy preferred outcome in finitely repeated coordination games. Crawford and Haller [6] study learning in repeated coordination games when the description of the game is not common knowledge. Fudenberg and Levine [8] show that in a repeated game of incomplete information with one long-lived player against a series of short-lived players, the long-lived player attains her Stackelberg payoff.

[5] Proposition 1 below is similar to Proposition 1 in Farrell [7], but in general the two yield different results (e.g., in Fig. 6 of Ben-Porath and Dekel [3]).

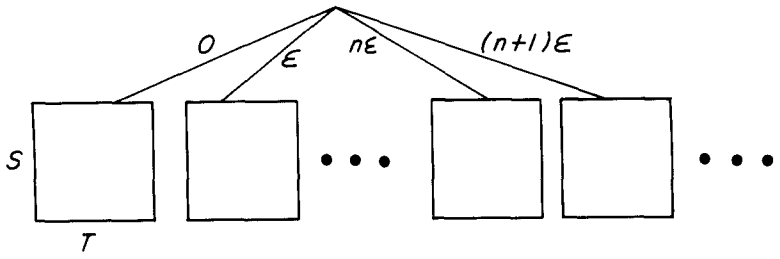

FIGURE 2.1

utility function, $u: S \times T \to \mathfrak{R}$, to mixed strategies, and similarly v is also extended to mixed strategies.

The solution concept we apply to games is iterated deletion of weakly dominated strategies.[6] In a game $\langle S, T, u, v \rangle$, the strategy $s \in S$ is weakly dominated if there exists $\sigma \in \Delta(S)$ such that $u(\sigma, t) \geq u(s, t)$ for all $t \in T$, and $u(\sigma, \bar{t}) > u(s, \bar{t})$ for some $\bar{t} \in T$. We allow for any order of deletion that is maximal at each stage, where by maximal we mean that, if at any stage in the iteration any strategy of a player is deleted, then *all* the strategies of that player which are weakly dominated at that stage are deleted. (Thus, at any stage one can delete all weakly dominated strategies for either player, or for both.) The following proposition is the generalization of the example in the introduction.

PROPOSITION 1. *Let G be a game in which there exists an s^* in S, and a t^* in T such that $u(s^*, t^*) > u(s, t)$ for all $(s, t) \neq (s^*, t^*)$ and $v(s^*, t^*) > v(s^*, t)$ for all $t \neq t^*$. Then there exists a $\delta > 0$ such that for any $\delta > \varepsilon > 0$ there is a unique outcome of $G(\varepsilon)$ which survives iterative maximal deletion of weakly dominated strategies. In this outcome $(n, s) = (0, s^*)$ and $\mathbf{t}(0) = t^*$ so the utilities are $u(s^*, t^*)$, $v(s^*, t^*)$.*

We will prove Proposition 1 for the case where at each stage all weakly dominated strageties of both players are deleted. A trivial modification of the argument proves the result for the more general case where at each stage all the strategies of one or both players are deleted. The proof involves several lemmas.

The following notation will be helpful:

[6] Proposition 1 below is also true when iterated deletion of weakly dominated strategies is replaced by stability (Kohlberg and Mertens [12]). This follows from Kohlberg and Mertens ([12] Proposition 6) and the easily verifiable fact that in the games we consider the outcome which survives iterated deletion of weakly dominated strategies is a stable outcome.

$$u^* = u(s^*, t^*); \qquad v^* = v(s^*, t^*);$$

$$\sigma_* = \operatorname*{argmax}_{\sigma \in \Delta(S)} \min_{\tau \in \Delta(T)} u(\sigma, \tau); \qquad u_* = \max_{\sigma \in \Delta(S)} \min_{\tau \in \Delta(T)} u(\sigma, \tau);$$

$S^j(n) \equiv \{s \in S : (n, s) \text{ remains after } j-1 \text{ stages of deletion}\}$;

$N^j \equiv \{n \in N : S^j(n) \neq \varnothing\}$;

$\mathbf{T}^j \equiv \{\mathbf{t} \in \mathbf{T} : \mathbf{t} \text{ remains after } j-1 \text{ stages of deletion}\}$; and

$T^j(n) \equiv \{t \in T : \mathbf{t}(n) = t \text{ for some } \mathbf{t} \in \mathbf{T}^j\}$.

Now, set $\delta = \min\{u^* - u(s, t) : (s, t) \in S \times T,\ (s, t) \neq (s^*, t^*)\}$ and choose an ε such that $0 < \varepsilon < \delta$.

LEMMA 1. $\mathbf{T}^j = \prod_{n=0}^{\infty} T^j(n)$.

Proof. We show by induction on j that: (1) $\mathbf{T}^j = \prod_{n=0}^{\infty} T^j(n)$; and (2) a strategy $\mathbf{t}$ is weakly dominated at stage j if and only if for some $n \in N^j$, $\mathbf{t}(n)$ is weakly dominated in the game $\langle S^j(n), T^j(n), u, v\rangle$.

First we note that if (1) and (2) are satisfied for $j-1$ then (1) is satisfied for j. This follows because $\mathbf{T}^j$ is obtained from $\mathbf{T}^{j-1}$ by deleting all the strategies $\mathbf{t}$ with the property that, for some $n \in N^{j-1}$, $\mathbf{t}(n)$ is weakly dominated in the game $\langle S^{j-1}(n), T^{j-1}(n), u, v\rangle$.

We now show that if (1) and (2) are satisfied for $j-1$ then (2) is satisfied for j. First suppose that, at stage j, the strategy $\mathbf{t}$ is weakly dominated by $\tau = (\tau(1), \tau(2), \ldots)$. Then there exists a strategy $(n, s') \in N^j \times S^j(n)$ such that $\bar{v}((n, s'), \mathbf{t}) < \bar{v}((n, s'), \tau)$. So, $v(s', \mathbf{t}(n)) < v(s', \tau(n))$. Also, $v(s, \mathbf{t}(n)) \leqslant v(s, \tau(n))\ \forall s \in S^j(n)$. Thus, $\mathbf{t}(n)$ is weakly dominated by $\tau(n)$. Next, consider the converse. Given a $\mathbf{t} \in \mathbf{T}^j$, suppose that $\mathbf{t}(n) \in T^j(n)$ is weakly dominated by $\tau \in \Delta[T^j(n)]$, for some $n \in N^j$. Clearly $\mathbf{t}$ is weakly dominated by τ where τ is defined as follows: $\tau(n) = \tau$ and, $\forall n' \neq n$, $\tau(n') = \mathbf{t}(n')$. It remains to show that $\tau \in \Delta(\mathbf{T}^j)$. This follows because τ is a probability distribution over strategies $\mathbf{t}_1, \ldots, \mathbf{t}_m$ such that $\mathbf{t}_i(k) \in \mathbf{T}^j(k)$, for $i = 1, \ldots, m$. By the induction hypothesis and the claim in the preceding paragraph, $\mathbf{t}_i \in \mathbf{T}^j$, for each $i = 1, \ldots, m$.

LEMMA 2. *If* $(n, s^*) \in N^j \times S^j(n)$ *then* $t^* \in T^j(n)$.

Proof. If $t^* \notin T^j(n)$ then, by Lemma 1, at some stage $i < j$, t^* was weakly dominated by some $\tau \in \Delta[T^i(n)]$ in the game $\langle S^i(n), T^i(n), u, v\rangle$. But this contradicts $v(s^*, t^*) > v(s^*, t)\forall t \neq t^*$.

LEMMA 3. *If* (n, s^*) *is deleted at stage* j *then there exists a strategy* $(\eta, \boldsymbol{\sigma})$ *such that* $\eta(n) = 0$ *and* $\bar{u}((\eta, \boldsymbol{\sigma}), \mathbf{t}) \geqslant u^* - n\varepsilon$, $\forall \mathbf{t} \in \mathbf{T}^j$.

Proof. Let (η, σ) be the strategy which weakly dominates (n, s^*). Assume first that $\eta(n)=0$. By Lemma 2 there exists $\mathbf{t}' \in \mathbf{T}^j$ such that $\mathbf{t}'(n)=t^*$. By Lemma 1, given any $\mathbf{t} \in \mathbf{T}^j$ there exists a $\mathbf{t}'' \in \mathbf{T}^j$ such that $\mathbf{t}''(n')=\mathbf{t}(n') \forall n' \neq n$ and $\mathbf{t}''(n)=t^*$. Since $\eta(n)=0$, $\bar{u}((\eta, \sigma), \mathbf{t})=\bar{u}((\eta, \sigma), \mathbf{t}'') \geqslant \bar{u}((n, s^*), \mathbf{t}'')=u^*-n\varepsilon$. If $\eta(n) \neq 0$ consider the strategy (η', σ) where $\eta'(n')=(\eta(n'))/(1-\eta(n))$ for all $n' \neq n$ and $\eta'(n)=0$. Consider any $\mathbf{t} \in \mathbf{T}^j$ such that $\mathbf{t}(n)=t^*$. Since $u(\sigma(n), t^*) \leqslant u(s^*, t^*)$ and $\bar{u}((\eta, \sigma), \mathbf{t}) \geqslant u^*-n\varepsilon$ it follows that $\bar{u}((\eta', \sigma), \mathbf{t}) \geqslant u^*-n\varepsilon$. For a $\mathbf{t} \in \mathbf{T}^j$ with $\mathbf{t}(n) \neq t^*$ an argument similar to that above can be used to construct a $\mathbf{t}''$ and show that $\bar{u}((\eta', \sigma), \mathbf{t}) \geqslant u^*-n\varepsilon$.

LEMMA 4. *If $(m+1, s^*)$ is deleted at stage j, then (m, s), for any $s \neq s^*$, is also deleted at stage j.*

Proof. It follows from the definition of δ and the choice of ε that $[u^*-(m+1)\varepsilon]-[u(s, t)-m\varepsilon] \geqslant \delta-\varepsilon>0$, for all $s \neq s^*$ and all t. By Lemma 3, there exists at stage j a strategy $(\eta, \bar{\sigma})$ such that $\bar{u}((\eta, \bar{\sigma}), \mathbf{t}) \geqslant u^*-(m+1)\varepsilon$, for all $\mathbf{t} \in \mathbf{T}^j$. It follows that, at stage j, $(\eta, \bar{\sigma})$ (strongly) dominates (m, s), for all $s \neq s^*$. Therefore, for all $s \neq s^*$, the strategy (m, s) is deleted.

Proof of Proposition 1. Assume that the iteration process has been completed at stage i so that $N^i=N^{i+1}$, $S^i=S^{i+1}$, and $\mathbf{T}^i=\mathbf{T}^{i+1}$. Let $m=\max\{n: n \in N^i\}$. (Note that m is finite because if k is an integer such that $k\varepsilon>u^*-u_*$ then, for all s, (k, s) is strongly dominated by $(0, \sigma_*)$. Therefore (k, s) is deleted at the first stage of the iteration.) Let $j<i$ be the stage at which $(m+1, s^*)$ was deleted. By Lemma 4, all (m, s) such that $s \neq s^*$ are deleted at stage j as well. So, $S^j(m)=\{s^*\}$. Therefore $\mathbf{T}^i(m)=\{t^*\}$. If $m=0$ the proof is complete. If not, then, as above, the strategies $(m-1, s)$ for all $s \neq s^*$, and all the strategies $\mathbf{t} \in \mathbf{T}^j$ with $\mathbf{t}(m-1) \neq t^*$, have been deleted. Therefore, at some stage $l<i$, $(m-1, s^*)$ dominates (m, s^*). But this contradicts our assumption that $m \in N^i$. Q.E.D.

Next, several issues regarding the specification of the game and the signaling stage are discussed. After two remarks, the importance of the order in which players can signal is discussed. Henceforth, the pure strategies of player 1 in the extension of a game G are denoted by (b, s)

	L	R
U	90,90	0,0
D	0,0	72,72

FIGURE 2.2a

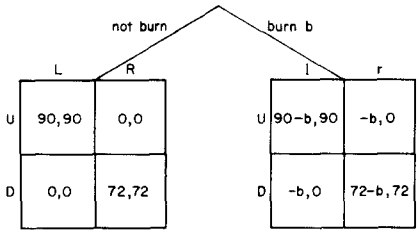

FIGURE 2.2b

where $b \in \mathfrak{R}_+$ is the amount burned and $s \in S$ is the strategy played after burning b. (This should not be confused with the previous notation (n, s) which indicates player 1 burning $n\varepsilon$ as discussed above.)

Remark 2.1. In the example in the introduction, player 1 only needed the option of burning one particular sum in order to achieve the preferred outcome, rather than using several levels of burning as in the proof. To see that the latter is in fact necessary, even in a simple example, consider the game in Fig. 2.2a and its extension in Fig. 2.2b. For any value of b the procedure of iteratively deleting weakly dominated strategies will not delete the strategies $(0, U)$ and $(0, D)$ of player 1, nor will it delete the strategies of player 2 which specify playing L and R after observing 0. To see this note that the maximin payoff for player 1 is 40. Therefore (b, D) is weakly dominated if and only if $b \geqslant 32$. However, since $90 - 32 < 72$ the iteration ends at this stage. Hence, for no b could player 1 guarantee her preferred outcome by having the option of burning b. Proposition 1 implies that if other amounts are available, then $((0, U); L)$ will be the outcome.

Remark 2.2. Proposition 1 also holds if player 1 has a continuum of burning options (not only discrete amounts). The proof follows similar lines. On the other hand, if the players do not have finitely many strategies in the game G, then the signaling may fail. To see this consider an example where $S = T = [0, 1]$, $u(s, t) = v(s, t) = s$ if $s = t$, and $u(s, t) = v(s, t) = 0$ otherwise. This game has a continuum of Nash equilibria: $\{(s, t) : s = t\}$. In the extension of this game *all* these outcomes remain. (After burning any amount $b \in [0, 1]$ player 1 will not play $s \leqslant b$ since then (b, s) is weakly dominated by $(0, 0)$. This only implies that 2 will play some $\mathbf{t}(b) > b$ and the iteration ends at this stage.) Nevertheless, in infinite games players can signal their future strategy if the outcome in the infinite game is defined as the limit of the outcomes in a sequence of finite approximations.[7]

A special class of games which satisfy the hypothesis of Proposition 1 are those where player 1's opponent also prefers the outcome (s^*, t^*) to all

[7] For details see Ben-Porath and Dekel [3].

a

	L	R
U	9,9	0,7
D	7,0	6,6

FIGURE 2.3a

other outcomes. In these games, where the players' interests coincide, signaling by either player will yield the mutually preferred outcome. Moreover the result that signaling selects the mutually preferred outcome can be extended to n-person games of this form (i.e., with an outcome which is preferred by all the players) but only under a specific assumption about the order in which the signaling occurs. The assumption is that: (i) players signal in sequence; (ii) each player "leaves the scene after she signals," so that she observes only the signals of players who precede her in the sequence; and finally (iii) all the players choose their actions in G simultaneously.[8]

Consider now the extension of the game G in Fig. 2.3a, when both players first burn simultaneously, and then after they each observe how much the other has burned, they both chose their actions in G (simultaneously). We assume that the players can only burn one amount, namely 1.5.[9] The normal form for this extension of G is given in Fig. 2.3b, where a strategy is a triplet whose first letter indicates whether the player burns (B) or not (0); the second letter indicates what action in G the player chooses if the opponent did not burn; and the third lettter indicates what action in G is chosen if the opponent was observed to burn. In this game only bDD and bRR are weakly dominated, so simultaneous signaling need not select the mutually preferred outcome.[10] In particular, both players burning and then not cooperating is an outcome that survives iterated deletion of weakly dominated strategies. Intuitively, this outcome arises when each player thinks that cooperation will follow if and only if she burns and the other does not burn.

The preceding two paragraphs demonstrated that the timing of the signaling is crucial even in games where the players' interests coincide. This emphasizes a problem that occurs in applying signaling: how does a player obtain the right to burn? When the players' objectives differ they will both want to burn. Thus, if players signal one after the other and then play the

[8] For details see Ben-Porath and Dekel [3].

[9] The conclusion below holds even if players are allowed to burn arbitrary amounts.

[10] Van Damme [18] makes a similar point.

b	*OLL*	*OLR*	*ORL*	*ORR*	*bLL*	*bLR*	*bRL*	*bRR*
OUU	9,9	9,9	0,7	0,7	9, 7.5	9, 7.5	0, 5.5	0, 5.5
OUD	9,9	9,9	0,7	0,7	7,-1.5	7,-1.5	6, 4.5	6, 4.5
ODU	7,0	7,0	6,6	6,6	9, 7.5	9, 7.5	0, 5.5	0, 5.5
ODD	7,0	7,0	6,6	6,6	7,-1.5	7,-1.5	6, 4.5	6, 4.5
BUU	7.5, 9	-1.5, 7	7.5, 9	-1.5, 7	7.5, 7.5	-1.5, 5.5	7.5, 7.5	-1.5, 5.5
BUD	7.5, 9	-1.5, 7	7.5, 9	-1.5, 7	5.5, -1.5	4.5, 4.5	5.5, -1.5	4.5, 4.5
BDU	5.5, 0	4.5, 6	5.5, 0	4.5, 6	7.5, 7.5	-1.5, 5.5	7.5, 7.5	-1.5, 5.5
BDD	5.5, 0	4.5, 6	5.5, 0	4.5, 6	5.5, -1.5	4.5, 4.5	5.5, -1.5	4.5, 4.5

FIGURE 2.3b

game in Fig. 1.1, one might ask which player is more "powerful," the first or the second to burn? It is easy to verify that the player who can signal last is most powerful. This can be seen by noting that in any subgame following an amount burned by the first player, the second can achieve her best outcome. The first player cannot convincingly communicate her intent to play the strategy which leads to her preferred outcome.

In the last two games discussed above a player was unable to signal credibly because after burning she could condition her choice of action in the game on whether another player burned or not. Thus, the ability of the

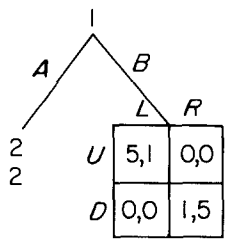

FIGURE 2.4

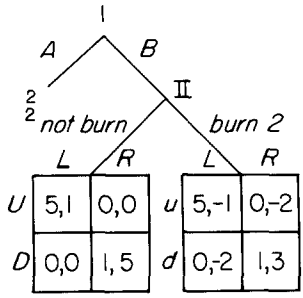

FIGURE 2.5

receiver to counter-signal can invalidate the credibility of the sender's signal. It is worth emphasizing that *any* signaling of future actions, and not only signaling by having the option to incur costs, is vulnerable to counter-signaling. In particular, signaling by forgoing an option can also be invalidated by counter-signaling. To see this, consider the game in Fig. 2.4 (due to Kohlberg and Mertens [12]). In this game iterated elimination of dominated strategies implies that player 1 plays B and then players 1 and 2 coordinate on U, L. Intuitively, by giving up 2, player 1 credibly signals that she will subsequently play U. Assume now that if player 1 plays B, then player 2 has the option to burn 2. In this game, see Fig. 2.5, iterated elimination of dominated strategies selects the outcome where player 1 plays A. Player 1 choosing B and forgoing 2 is no longer a credible signal because player 2 has the option to counter-signal by burning 2.

This vulnerability to counter-signals distinguishes signaling future actions from signaling types. Obviously, when a player signals private information about something not under her control (as, for example, in Cho and Kreps [5]) there is no point for the receiver to object.

3. CONCLUSION

We have shown that forwards induction implies that in a class of games the potential of incurring a cost signals a future action. However, when all players can incur costs the receiver can "object" to a signal of a future action by counter-signaling. This vulnerability to counter-signaling distinguishes signaling future actions from signaling types, and suggests that in reality signaling future actions is problematic.[11]

[11] Rubinstein [16] makes a related point: he observes that people do not perceive some forms of costly signaling, such as burning money, as relevant to the game, and he proposes properties of signals that may be useful in determining whether they are relevant.

APPENDIX

This appendix provides an example based on two standard models: the Spence [17] signaling model and the Nash [14] bargaining game. The example shows how models that focus on signaling types (such as Spence [17], Milgrom and Roberts [13], and Cho and Kreps [5]) can be naturally extended so that the option of incurring a cost enables the player not only to signal his type but also (at the same time) to signal a future action. The example also shows that in many environments "burning" is a naturally available strategy—all signaling games have such strategies.

The intuition for the example is based on Proposition 1. In a simple signaling model player 1 may be one of several types, and she chooses a costly signal. This signal is observed by player 2, who then chooses an action. The type–signal–action triplet determines the payoffs for each player. Consider now replacing player 2's choice of action by a more general (simultaneous move) game between players 1 and 2, denoted G. The costly signaling can now be viewed as a "burning" stage—any type of player 1 could always "burn" by selecting a costlier signal. Proposition 1 suggests that if there are multiple equilibria in G, then player 1's ability to burn enables her to influence which of these equilibria will be played.

The model analyzed is a modification of a stylized version of Spence's signaling model. In the stylized version (see also Cho and Kreps [5]) there is a worker who may be of high (h) or low (l) quality. This worker moves first, choosing an education level b in $[0, \infty)$. The worker is paid a wage, w in W, by a firm. (How the wage is determined is described below.) The worker's worth to the firm is \$2 if he is h and \$1 if l. The worker's utility is $w - b$ if he is l, and $w - b/2$ if he is h. In the unmodified model the firm pays the worker his expected value, where the firm's beliefs are determined in an equilibrium by: (i) her prior over $\{l, h\}$; (ii) the worker's equilibrium strategy; and (iii) the observed signal b. (The specification that the firm pays the worker his expected value can be justified by a model where there are several firms competitively bidding for the worker in the style of Bertrand.) Our modification is to assume that the firm and worker determine the wage by a (modification of the) Nash [14] non-cooperative bargaining game. That is, the firm and worker both simultaneously announce a wage. If the firm's offer is at least as large as the worker's request then the worker is paid his request and works for the firm. Otherwise the worker is not employed. In the latter case the firm's utility is zero and worker's utility is $-b$. If the worker is employed then the worker's utility is as described above ($w - b$, or $w - b/2$) and the firm earns $1 - w$ or $2 - w$, depending on whether the worker is l or h, respectively. Formally, we denote the (pure) strategies for the worker by a quadruplet

$[(b_l, w_l), (b_h, w_h)]$ which indicates the education the worker purchases and the wage he requests, if he is a low or high type, respectively. The firm's strategy is a function $S: [0, \infty) \to W$ determining the wage paid as a function of the education observed.

Proposition 1 suggests that the worker can use the level of education to signal his wage demand in the bargaining game, as well as his type. This is demonstrated below by using a refinement of Nash equilibrium which is motivated by forwards-induction rationality. The refinement is a version of the intuitive criterion (Cho and Kreps [5]), and it tests the plausibility of equilibrium outcomes. An equilibrium outcome $\mathscr{E}$ is specified by the equilibrium path, $\mathscr{E} = \{[(b_l, w_l), (b_h, w_h)], [S(b_l), S(b_h)]\}$. (We consider here only pure strategy equilibria; however, the arguments below can be extended to incorporate mixed strategies without effecting the conclusion.) A strategy of the worker fails the criterion if, regardless of the firm's strategy, and for $\alpha = h$ or $\alpha = l$, the level of education and wage request of an α quality worker in this strategy lead to a lower payoff than the α type receives in $\mathscr{E}$. Formally, given the outcome $\mathscr{E}$, a strategy $[(\tilde{b}_l, \tilde{w}_l), (\tilde{b}_h, \tilde{w}_h)]$ fails the criterion if $\tilde{w}_l - \tilde{b}_l < w_l - b_l$ or $\tilde{w}_h - \tilde{b}_h/2 < w_h - b_h/2$. Strategies which are either weakly dominated or fail this criterion are iteratively deleted. If in the game which remains after these deletions there is no Nash equilibrium which leads to the same outcome as $\mathscr{E}$, then $\mathscr{E}$ is said to fail the test. It is important, as in the previous sections, that W be discrete. For simplicity assume that $W = \{k/10 : k \in N,\ k \neq 10,\ k \neq 20\}$. As in Cho and Kreps [5], any outcome which fails such a test is not a stable outcome (in a game where the level of education, as well as W, is restricted to a finite set, since stable outcomes are defined for finite games). We prove below that the unique outcome which survives the test is $\{[(0, 0{,}9), (1, 1.9)], [0.9, 1.9]\}$.

The conclusion is then that in the unique outcome which survives the test the worker signals his type and receives his best payoff in the bargaining subgame. Note that the method in which the worker's wage is determined in the subgame was modified from that in the original signaling game: instead of assuming that the firm pays the worker all the expected surplus, the worker and firm must bargain over this surplus. Despite this modification the unique equilibrium outcome which survives the test is the same. This is because the signaling power which is granted to the worker in order to signal his type, also gives the worker the power to signal his wage request. The firm's best response is then to accept the worker's demands rather than lose the worker.

We now prove that the unique outcome which survives the test is $\mathscr{E}^* \equiv \{[(0, 0.9), (1, 1.9)], [0.9, 1.9]\}$. The proof is made up of several claims. Let $\mathscr{E} = \{[(b_l, w_l), (b_h, w_h)], [S(b_l), S(b_h)]\}$ be an equilibrium outcome which survives the test.

CLAIM 1. *$\mathscr{E}$ is not a pooling outcome, i.e., $b_l \neq b_h$.*[12]

Proof. Assume by contradiction that $\mathscr{E}$ is a pooling outcome, and let w denote the wage in $\mathscr{E}$. The employer is not making negative expected payoffs, and therefore $w \leqslant p_l + 2p_h$.

Let w' be the largest element in W that is smaller than 2. By assumption, $w < w'$. It is easy to see that there exists a level of education, b, such that $w' - b/2 > w - b_h/2$, $w' - b < w - b_l$, and for every $\tilde{w} < w'$ in W we have $\tilde{w} - b/2 < w - b_h/2$ (where $b_l = b_h$ is the education level in $\mathscr{E}$).

Clearly all the strategies where the employer makes an offer that is larger than 2 are weakly dominated and can be deleted. Next, all the strategies of the worker where the low type chooses an education level b or the high type chooses b and a wage that is smaller than w' fail the criterion and are deleted. At the next stage of the iteration all the strategies of the employer which offer less than w' in response to b are weakly dominated. (Because, if the worker chooses b he must be the high type, and therefore must be asking for w' or more. If the employer offers less than w' the worker will not be employed.) But now, in the game that remains after these deletions, there is no Nash equilibrium with the outcome $\mathscr{E}$, since the high type can benefit by deviating to education level b and wage demand w'. Q.E.D.

CLAIM 2. *The low quality worker receives the highest wage which is less than his value, i.e. $w_l = 0.9$.*

Proof. Assume not, so that in $\mathscr{E}$ $w_l < 0.9$. Let $\tilde{b}_l$ satisfy $b_l < \tilde{b}_l < b_l + 0.1$. Then all the worker's strategies where either type gets an education level of $\tilde{b}_l$ and requests a wage less than or equal to w_l fail the criterion and are deleted. Hence (in the next stage of the iteration) any strategy which offers w_l or less to a worker with education $\tilde{b}_l$ is weakly dominated. This is because offering w_l or less will surely lead to the worker being unemployed, while offering $w_l + 0.1$ leads to a positive profit if the worker is employed (since $w_l + 0.1 < 1$). Hence in the game which remains after these deletions, there is no Nash equilibrium leading to the outcome $\mathscr{E}$ (because the worker will prefer to deviate to $[(\tilde{b}_l, w_l + 0.1), (b_h, w_h)]$). Q.E.D.

CLAIM 3. *The high quality worker receives the highest wage which is less than his value, i.e., $w_h = 1.9$.*

Proof. All the strategies of the firm which offer wages greater than 1.9, for any b, are weakly dominated. Assume that in $\mathscr{E}$ $w_h < 1.9$. Then there exists $\tilde{b}_h$ such that $w_l - b_l > 1.9 - \tilde{b}_h$, $w_h - b_h/2 < 1.9 - \tilde{b}_h/2$, and

[12] In order to avoid a pooling equilibrium, W must be chosen to be sufficiently fine. In particular, the largest element of W which is strictly less than 2 must be greater than $p_l + 2p_h$ (where p_α, $\alpha = l, h$, is the prior probability of a worker of type α).

$w_h - b_h/2 > 1.8 - \tilde{b}_h/2$. Then all the strategies of the worker where the low type gets an education level of $\tilde{b}_h$ are deleted. Also, all the worker's strategies where the high type gets an education level of $\tilde{b}_h$ and requests less than 1.9 are deleted. So, as in the proof of Claim 2, any strategy for the firm which offers less than 1.9 in response to $\tilde{b}_h$, is weakly dominated. Hence, in the game which remains after these deletions, the outcome $\mathscr{E}$ is not an equilibrium outcome (because the worker will deviate to $[(b_l, w_l), (\tilde{b}_h, 1.9)]$). Q.E.D.

CLAIM 4. $b_l = 0$.

Proof. Assume that in $\mathscr{E}$, $b_l > 0$, and let $\tilde{b}_l$ satisfy $b_l - 0.1 < \tilde{b}_l < b_l$ (and $\tilde{b}_l > 0$). Clearly any strategy of the worker in which one of the types chooses the education level of $\tilde{b}_l$ and requests $\tilde{w}_l$ where $\tilde{w}_l < w_l$, can be deleted. So in response to $\tilde{b}_l$ the firm should offer at least w_l. Then, in the game remaining after these deletions, the worker will deviate from any strategy which supports the outcome $\mathscr{E}$. Q.E.D.

CLAIM 5. $b_h = 1$.

Proof. Since $w_h = 1.9$, $b_l = 0$, and $w_l = 0.9$ it must be the case that $b_h \geqslant 1$ (otherwise the low type will imitate the high quality worker). So assume that $b_h > 1$. Let $\tilde{b}_h$ satisfy $b_h - 0.1 < \tilde{b}_h < b_h$ (and $1 < \tilde{b}_h$). As before (see Claim 3) any strategy of the worker where either: (a) the low type chooses an education level of $\tilde{b}_h$; or (b) the high quality worker chooses $\tilde{b}_h$ and requests less than 1.9, can be deleted. Hence, in response to $\tilde{b}_h$, the firm should offer w_h. Again, this implies that an outcome where $b_h \neq 1$ fails the test. Q.E.D.

REFERENCES

1. R. J. AUMANN AND S. SORIN, Cooperation and bounded rationality, *Games Econ. Behavior* **1** (1989), 5–39.
2. J. S. BANKS AND J. SOBEL, Equilibrium selection in signalling games, *Econometrica* **55** (1987), 647–662.
3. E. BEN-PORATH AND E. DEKEL, Coordination and the potential for self sacrifice, mimeo, Granduate School of Business, Stanford University, Nov. 1987.
4. I.-K. CHO, A refinement of sequential equilibrium, *Econometrica* **55** (1987), 1367–1390.
5. I.-K. CHO AND D. M. KREPS, Signalling games and stable equilibria, *Quart. J. Econ.* **CII** (1987), 179–221.
6. V. P. CRAWFORD AND H. HALLER, Learning how to cooperate: Optimal play in repeated coordination games, *Econometrica* **58** (1990), 571–595.
7. J. R. FARRELL, Communication and Nash equilibrium, *Econ. Lett.* **27** (1988), 209–214.
8. D. FUDENBERG AND D. LEVINE, Reputation and equilibrium selection in games with a patient player, *Econometrica* **54** (1989), 759–778.

9. J. GLAZER AND A. WEISS, Pricing and coordination: Strategically stable equilibria, *Games Econ. Behavior* **2** (1990), 118–128.
10. J. C. HARSANYI, Games of incomplete information played by Bayesian players, I, II, and III, *Manage. Science.* **14** (1967–1968), 159–182, 320–334, 486–502.
11. E. KALAI AND D. SAMET, Unanimity games and Pareto optimality, *Int. J. Game Theory* **14** (1985), 41–50.
12. E. KOHLBERG AND J.-F. MERTENS, On the strategic stability of equilibria, *Econometrica* **54** (1986), 1003–1038.
13. P. MILGROM AND J. ROBERTS, Price and advertising signals of product quality, *J. Polit. Econ.* **XCIV** (1986), 796–821.
14. J. NASH, Two-person cooperative games, *Econometrica* **21** (1953), 129–140.
15. M. J. OSBORNE, Signalling, forward induction, and stability in finitely repeated games, *J. Econ. Theory* **50** (1990), 22–36.
16. A. RUBINSTEIN, "Comments on the Interpretation of Game Theory," Discussion Paper No. TE/88/181 STICERD, London School of Economics, 1988, forthcoming in *Econometrica.*
17. M. A. SPENCE, "Market Signalling," Harvard Univ. Press, Cambridge, MA, 1974.
18. E. VAN DAMME, Stable equilibria and forward induction, *J. Econ. Theory* **48** (1989), 476–496.