JOURNAL OF ECONOMIC THEORY 21, 1-9 (1979)

Equilibrium in Supergames with the Overtaking Criterion*

ARIEL RUBINSTEIN

Department of Economics, The Hebrew University, Jerusalem, Israel

Received October 17, 1977; revised September 14, 1978

1. INTRODUCTION

There are significant differences between the situation of players undertaking to play a single game, and players who know that they will play the same game repeatedly in the future. Strategy in the first case is a single play; in the second, it is a sequence of rules, each one of which designates the play at the corresponding game and may pertain to the outcomes preceding. The preferences of the participants are determined partly by temporal considerations, and the participants may adopt "risky" strategies, "protected" by threats of retribution in the future.

A finite sequence of identical games is an inadequate model for examining the idea of repeated games, as is shown by the following analysis. (For a detailed analysis, see [11]). If the number of games is finite and known initially, the players will treat the last game as if it were a single game. Thus the threats implicit in the game before last are proven to be false threats. Therefore the game before last will be treated as a single game, and so on. Thus the situation we wish to describe is not expressed by such a sequence.

In order to avoid "end-points" in the model, we define a supergame. A supergame is an infinite sequence of identical games, together with the players' evaluation relations (that is their preference orders on utility sequences). Obviously the assumption of an infinite planning horizon is unrealistic, but it is an approximation to the situation we wish to describe.

The literature mainly compares the equilibrium concepts in supergames and single games (see Aumann [1-5]). Other papers emphasize the uses of the concept of supergames in economies ([8-10]).

In most of the papers, it was assumed that the participants evaluate the utility flows according to the criterion of the limit of the means of the flows (but see [8].) The drawback of this evaluation relation is that it ingores any finite time interval. The aim of this paper is to extend the discussion to supergames with evaluation relations determined according to the "overtaking

* I wish to thank Professor Peleg for his advice and guidance.

0022-0531/79/040001-09$02.00/0

Copyright © 1979 by Academic Press, Inc.
All rights of reproduction in any form reserved.

criterion". (The sequence $\{a_t\}_{t=1}^{\infty}$ is preferred to the sequence $\{b_t\}_{t=1}^{\infty}$ if $0 < \underline{\lim} \sum_{t=1}^{T} a_t - b_t$.)

The formal model, described in Section 2, is taken from Roth.[1] The single game is given in strategic form (see [5]).

Let n be the number of players in the game. A (Nash) equilibrium in the supergame is an n-tuple of supergame strategies such that no player on his own can deviate profitable from his strategy. A stationary equilibrium is an equilibrium which, if adhered to by all players, will produce identical outcomes for every game played. The stationary equilibria will be characterized in Section 3.

The "power" of the threats makes possible the existence of many equilibrium points, some of which may satisfy further requirements. An equilibrium will be called perfect if after any possible "history", the strategies planned are an equilibrium. In other words no player ever has a motivation to change his strategy. This will be treated in Section 4, where it is shown that the requirement of perfection alters the outcomes of stationary equilibria only "marginally".

The result corresponds to similar results obtained in [6] and [13] for the limit of the means criterion, and in [10] for a model of altruistic behavior. Together, these results give the impression that the concept of perfection does not enable the isolation of a small solution set, from the Nash equilibria. In [13, 14] an example is given to demonstrate that perfection is a significant notion considering strong equilibria in supergames where the evaluation relations are according to the overtaking criteria.

2. The Model

The single game G is a game in strategic form

$$G = \langle \{S_i\}_{i=1}^{n}, \{\pi_i\}_{i=1}^{n} \rangle.$$

The set of players is $N = \{1,..., n\}$. For each $i \in N$, the set of strategies of i is S_i; S_i is assumed non-empty and compact. Each player i has a payoff function $\pi_i : S \to \mathscr{R}$($\mathscr{R}$ — the reals), which is continuous in the product topology.

Given $\sigma \in S$, a payoff vector is the n-tuple $\pi(\sigma) = \langle \pi_1(\sigma),..., \pi_n(\sigma) \rangle$. For convenience we will denote the $n-1$-tuple $\langle \sigma_1,..., \sigma_{i-1}, \sigma_{i+1},..., \sigma_n \rangle$ by σ^{-i}, and the n-tuple σ by $\langle \sigma^{-i}, \sigma_i \rangle$. σ will be called a (Nash) equilibrium if for all i and for all $s_i \in S_i$, $\pi_i(\sigma^{-i}, s_i) \leqslant \pi_i(\sigma)$.

[1] I wish to thank Professor A. E. Roth for permitting me to use the model described in [12].

If the set of strategies is finite and it is possible to adopt mixed strategies, we can identify S_i with the set of mixed strategies, and π_i with the expected payoff of i. Examples in a similar context may be found in [8] and [9].

The supergame, G^∞ is $\langle G, \prec_1, ..., \prec_n\rangle$ where G is a single game and the $\prec_i$'s are evaluation relations on real number sequences; more exactly, $\prec_i$ is a binary relation on $\mathcal{R}^{\mathcal{N}}$[2] which is transitive, anti-symmetric, but not necessarily a total order.

The set of outcomes at time t, $S(t)$, is S. A strategy for i in G^∞ is a set $\{f_i(t)\}_{t=1}^\infty$, where $f_i(1) \in S_i$, and for $t \geqslant 2$, $f_i(t) : \prod_{j=1}^{t-1} S(j) \to S_i$. Thus a supergame strategy is a choice of strategies at every stage, where each choice is possibly dependent on the outcomes of the preceding games, and where all players know all the choices made by every player in the past.

The set of supergame strategies of i will be denoted by F_i. F is the set of n-tuples of strategies; $F = \prod_{i=1}^n F_i$.

Given $f \in F$, the outcome at time t will be denoted by $\sigma(f)(t)$, and is defined inductively by

$$\sigma(f)(1) = (f_1(1), ..., f_n(1))$$

$$\sigma(f)(t) = (..., f_i(t)(\sigma(f)(1), ..., \sigma(f)(t-1)), ...).$$

We will define a relation $\overline{\prec}_i$ on F, induced by $\prec_i$, as follows:

For all $f, g \in F, f \overline{\prec} g$ if and only if

$$\{\pi_i(\sigma(f)(t))\}_{t=1}^\infty \prec_i \{\pi_i(\sigma(g)(t))\}_{t=1}^\infty .$$

Given $f \in F$, we will denote $(f_1, ..., f_{i-1}, f_{i+1}, ..., f_n)$ by f^{-i}, and f by (f^{-i}, f_i).

$f \in F$ is stationary if there exists $\sigma \in S$ such that for all t, $\sigma(f)(t) = \sigma$. If $f \in F$ is stationary we will denote the corresponding σ by $\hat{\sigma}(f)$. Note that, in contrast with definitions appearing elsewhere in the literature, the stationary strategies produce constant outcomes (as in [12]).

$f \in F$ is a (Nash) equilibrium in the supergame G^∞ if for all i, and for all $h_i \in F_i$, $f \overline{\prec}_i (f^{-i}, h_i)$.

The main evaluation relation that was considered in the literature is the *Limit of means evaluation relation*, defined by

$$x \prec y \text{ iff } 0 < \underline{\lim}\left(\sum_{t=1}^{T} y_t - x_t\right)\Big/T.$$

[2] Let A be a set. $A^{\mathcal{N}}$ is the set of sequences of elements in A.

In the following we will concentrate on the *Overtaking criterion evaluation relation*, defined by

$$x \prec y \text{ iff } 0 < \underline{\lim} \sum_{t=1}^{T} (y_t - x_t).$$

Remark. An axiomatic characterization of the Overtaking Criterion is given in Brock [7].

There exists no utility function representing the overtaking criterion, that is, no function $u : \mathscr{R}^{\mathscr{N}} \to \mathscr{R}$ satisfies $u(x) < u(y) \leftrightarrow x \prec y$ for all x, y which are $\prec$ related.

For every $a \in R$, $(a, a, \ldots) \prec (a + 1, a, \ldots)$, and for every $a < b$, $(a + 1, a, \ldots) \prec (b, b, \ldots)$. Thus $\mathscr{R}^{\mathscr{N}}$ has $\aleph \prec$ segments which are disjoint and non-empty, while $(\mathscr{R}, <)$ has less.

3. Characterization of Stationary Equilibria

We will denote $v_i = \min_{r \in S} \max_{t_i \in S_i} \pi_i(r^{-i}, t_i)$. v_i is the minimal utility that the players apart from i may force on i.

DEFINITION. $s \in S$ is a weakly forced outcome in G if, for all i, $v_i \leqslant \pi_i(s)$.

Thus in a weakly forced outcome each player's payoff is at least as large as the amount the other players can force on him. The notion of payoff of weakly forced outcome is equivalent to the term individually rational payoff used in the literature.

DEFINITION. $s \in S$ is a strongly forced outcome in G if for all i, and for all $t_i \in S_i$, $\pi_i(s^{-i}, t_i) \leqslant \pi_i(s)$, or $v_i < \pi_i(s)$.

Thus a strongly forced outcome is an outcome in which any player who can gain from a deviation may be subject to a loss enforced on him by the other players. Any strongly forced outcome is of course a weakly forced outcome.

EXAMPLE. Let S_i be the set of mixed strategies of i, $i = 1, 2$ in a matrix game with payoff matrix

2, 2	0, 3
3, 0	1, 1

π_i is the expected payoff of i. (See Fig. 1.)

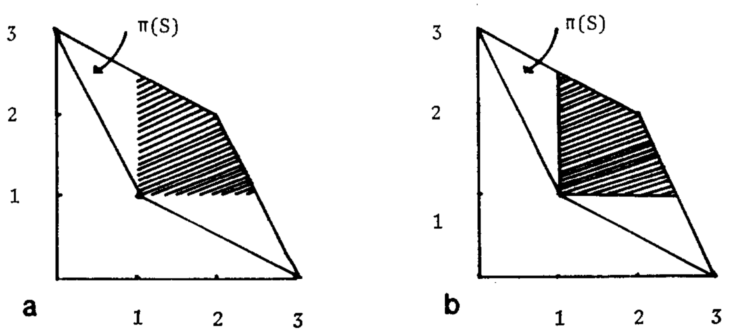

FIG. 1. (a) The payoff's set of the Strongly forced outcomes. (b) The payoff's set of the Weakly forced outcomes.

When the evaluation relations are according to the limit of the means, the set of outcomes which produce a stationary equilibrium in the supergame is the set of weakly forced outcomes (see [5], for example). Proposition 3.1 characterizes the stationary equilibria when the evaluation relations are according to the overtaking criterion.

PROPOSITION 3.1. *If for all i, $\prec_i$ is the overtaking evaluation relation, the stationary outcomes of equilibria are the strongly forced outcomes.*

Proof. Let $\sigma \in S$ be a strongly forced outcome. For any $i \in N$ define γ^i, the "punishing strategy" for i, by

(i) If $\pi_i(\sigma) = v_i$, $\gamma^i = \sigma$

(ii) Otherwise, let $\gamma^i \in S_i$ satisfy $\max_{s_i} \pi_i(\gamma^{i^{-i}}, s_i) < \pi_i(\sigma)$.

Define $f_i \in F_i$ for $i \in N$ as follows:

$$f_i(1) = \sigma_i$$

$$f_i(t)(s(1) \cdots s(t-1)) = \begin{cases} \gamma_i^j & \text{if there exists } T \leqslant t-1 \text{ such that } s(1) = \cdots = \\ & s(T-1) = \sigma \text{ and } s^{-j}(T) = \sigma^{-j} \text{ and } s_j(T) \neq \sigma_j. \\ \sigma_i & \text{otherwise.} \end{cases}$$

Then $\hat{\sigma}(f) = \sigma$ and it is obvious that f is a G^∞ equilibrium.

Let $f \in F$ be a stationary equilibrium. Denote $\hat{\sigma}(f) = \sigma$. Suppose that σ is not a strongly forced outcome. Then there exists $i \in N$ and $d_i \in S_i$ such that $\pi_i(\sigma^{-i}, d_i) > \pi_i(\sigma)$, and for any $r \in S$ there is $d_i(r) \in S_i$ such that $\pi_i(r^{-i}, d_i(r)) \geqslant \pi_i(\sigma)$. Now define $g_i \in F_i$

$$g_i(1) = d_i$$
$$g_i(t+1)(s(1) \cdots s(t)) = d_i(f(s(1) \cdots s(t))).$$

Clearly $f \prec_i (f^{-i}, g_i)$.

Remark. Consider the following payoff matrix of a two-players game:

1, 0	1, 0
2, 0	0, 2
0, 2	2, 0

As has previously been mentioned, by identifying mixed strategies and the expected payoff with S_i and π_i, matrix games are covered by the theory so far. The strategy (0, 1/2, 1/2) for the row player and (1/2, 1/2) for the column player is an equilibrium in the single game, and therefore the outcome is strongly forced. Thus it is a stationary outcome of an equilibrium in supergame, according to our definition of equilibrium.

Some alternative definitions are possible in the spirit of [1]. Thus a possible definition is an n-tuple supergame's strategies f is an equilibrium if there is no i and $\hat{f}_i$ such that i's payoff sequence resulting from adoption of $(f^{-i}, \hat{f}_i)$ if preferable with probability 1 to the expectation sequence resulting from adoption of f.

We will now show that ((0, 1/2, 1/2), (1/2, 1/2)) is not a stationary outcome of an equilibrium in this second definition. Let $f \in F$ satisfying $\hat{\sigma}(f) =$ ((0, 1/2, 1/2), (1/2, 1/2)).

The row player may deviate according to the following rule:

(i) he plays (0, 1/2, 1/2) until time T when the total payoff he has accumulated is $T + 1$,

(ii) he then deviates and plays (1, 0, 0), then the probability of such a T occurring, is 1. Since the row player's evaluation is according to the overtaking criterion, his payoff sequence will be preferable with probability 1 to that obtained had he not deviated.

4. Perfect Equilibria

The definition of equilibrium given in Section 2 was shown to be too general in Section 3. One possible restriction is that a deviation will prove unprofitable to a player at all stages of the game, and not only at the beginning.

Given $f \in F$ and $r(1),..., r(T) \in S$, the n-tuple of strategies determined by f after a "history" $r(1),..., r(T)$ is denoted by $f|_{r(1),...,r(T)}$; thus

$$(f|_{r(1),...,r(T)})_i\,(t)(s(1),..., s(t-1)) = f_i(T+t)(r(1),..., r(T), s(1),..., s(t-1)).$$

DEFINITION. $f \in F$ is a perfect equilibrium point if for all $r(1),..., r(T) \in S$, $f|_{r(1),...,r(T)}$ is an equilibrium.

Aumann and Shapley [6] and Rubinstein [13] proved that in a supergame with evaluation relations determined by the limit of the means criterion there is a perfect stationary equilibrium $f \in F$ such that $\hat{\sigma}(f) = \sigma$ iff σ is a weakly forced outcome.

But not every strongly forced outcome is the outcome of a perfect stationary equilibrium in a supergame, where all players have evaluation relations according to the overtaking criterion.

Consider the following matrix game where $S_1 = \{A_1, A_2\}$, and $S_2 = \{B_1, B_2, B_3\}$(mixed strategies are not allowed).

	B_1	B_2	B_3
A_1	(1, 1)	(0, 0)	(0, 0)
A_2	(2, 0)	(0, 0)	(2, 1)

$v_1 = 0$ and $v_2 = 1$.

(A_1, B_1) is a strongly forced outcome, but is not an outcome of a perfect stationary equilibrium; for if $(f_1, f_2) \in F$ is a perfect stationary equilibrium such that $\hat{\sigma}(f_1, f_2) = (A_1, B_1)$, then for all $s(1),..., s(t) \in S$,

$$f_2(t+1)(s(1),..., s(t)) = \begin{cases} B_1 \text{ if } f_1(t+1)(s(1),..., s(t)) = A_1 \\ B_3 \text{ if } f_1(t+1)(s(1),..., s(t)) = A_2 \end{cases}$$

But then the row player can profitably alter his strategy by $\hat{f}_1 \equiv A_2$, with a utility flow $(1, 1,...) \prec_2 (2, 2,...)$.

PROPOSITION 4.1. *In the supergame* $\langle G, \prec_1, ..., \prec_n \rangle$ *where* $\prec_i$ *are evaluation relations according to the overtaking criterion, and* $s \in S$ *satisfies* $v_i < \pi_i(s)$ *for all* i, *there exists* $f \in F$, *a stationary perfect equilibrium such that* $\hat{\sigma}(f) = s$.

Proof. The idea is to construct $f \in F$ such that a player deviating from the stationary position, or the punishing strategy of another player, will be punished sufficiently to eliminate his "profit". After punishment, the players return to the stationary position.

By assumption, there exist γ^i such that $\max_{t_i \in S_i} \pi_i(\gamma^{i^{-i}}, t_i) = \pi_i(\gamma^i) < \pi_i(s)$ (γ^i is strategy punishing i; the i'th component of γ^i is i's optimal defense strategy). We write $L_i = \pi_i(s) - \max_{t_i \in S_i} \pi_i(\gamma^{i^{-i}}, t_i) > 0$. L_i is the punishment i will receive every time the punishing strategy γ^i is employed against him. We will write $R_i = \max_{r \in \{s, \gamma^1, ..., \gamma^n\}} \{\max_{t_i \in S_i} \pi_i(t_i, r^{-i}) - \pi_i(r), \pi_i(s) -$

$\pi_i(r)\} \geqslant 0$. R_i is the maximal relative profit a player i can gain by deviating from one of the $n+1$ single game strategies deployed in f, and by bringing to an end the punishment of another player.

We will now define $m_i(s(1),\dots,s(t))$ and $f_i(t+1)(s(1),\dots,s(t))$ inductively as follows:

($m_i(s(1),\dots,s(t))$ is the length of time a player will be punished for participating in $s(1),\dots,s(t)$.)

$$m_i(\varnothing) = 0$$

$$f_i(1) = s_i .$$

$m_i(s(1),\dots,s(t+1)) =$

$$\begin{cases} \left[\frac{R_i}{L_i}\right] + 1 \text{ if for all } j,\ m_j(s(1),\dots,s(t)) = 0, s_i(t+1) \neq \sigma_i \text{ and } s^{-i}(t+1) = \sigma^{-i}. & (1) \\ \left[\frac{m_j(s(1),\dots,s(t)) \cdot R_i}{L_i}\right] + 1 \text{ if there exists } j \neq i \text{ such that } m_j(s(1),\dots, s(t)) > 0, s_i(t+1) \neq \gamma_i^{\,j} \text{ and } s^{-i}(t+1) = \gamma^{j^{-i}}. & (2) \\ m_i(s(1),\dots,s(t)) - 1 \text{ if } m_i(s(1),\dots,s(t)) > 0 \text{ and } s^{-i}(t+1) = \gamma^{i^{-i}}. & (3) \\ 0 \quad \text{otherwise.} & (4) \end{cases}$$

It is clear that for all $s(1),\dots,s(t) \in S$, the number of players i for whom $m_i(s(1),\dots,s(t)) > 0$ is at most 1.

Thus we can define

$$f_i(t+1)(s(1),\dots,s(t)) = \begin{cases} \gamma_i^{\,j} \text{ if } m_j(s(1),\dots,s(t)) > 0. \\ s_i \text{ otherwise.} \end{cases}$$

Clearly f is stationary and $\hat{\sigma}(f) = s$.

Let $r(1),\dots,r(T) \in S$, and let $\bar{f} = f|_{r(1),\dots,r(T)}$. We will show that $\bar{f}$ is an equilibrium.

Let $h_i \in F_i$; we will prove that $\bar{f} \not\prec_i (\bar{f}^{-i}, h_i)$. It is sufficient to show that if there exists t_0 such that $\pi_i(\sigma(\bar{f})(t_0)) < \pi_i(\sigma(\bar{f}^{-i}, h_i)(t_0))$, then there exists $t_0 < t_1$ such that $\sum_{t=t_0}^{t_1} \pi_i(\sigma(\bar{f})(t)) \geqslant \sum_{t=t_0}^{t_1} \pi_i(\sigma(\bar{f}^i, h_i)(t))$.

We will denote $m_i(\{\sigma(\bar{f}^{-i}, h_i)(t)\}_{t=1}^{t_0-1})$ by m_i. Player i cannot profitably deviate from γ^i since $\pi_i(\gamma_i) = \max_{t_i \in S_i} \pi_i(\gamma^{i^{-i}}, t_i)$. Therefore $m_i = 0$.

If $m_j = 0$ for all j, we will define $t_1 = t_0 + [R_i/L_i] + 1$; for all $t_0 < t \leqslant t_1$, $\sigma^{-i}(\bar{f}^{-i}, h_i) = \gamma^{i^{-i}}$, thus

$$\sum_{t=t_0}^{t_1} [\pi_i(\sigma(\bar{f}^{-i}, h_i)(t)) - \pi_i(\sigma(\bar{f}(t)))] \leqslant R_i - \left(\left[\frac{R_i}{L_i}\right] + 1\right) L_i \leqslant 0.$$

If $m_j > 0$, we will define $t_1 = t_0 + (m_j \cdot R_i/L_i) + 1$; for all $t_0 < t \leqslant t_1$, $\sigma^{-i}(f^{-i}, h_i) = \gamma^{i^{-i}}$, and thus

$$\sum_{t=t_0}^{t_1} \pi_i(s) - \pi_i(\sigma(\bar{f}(t))) + \pi_i(\sigma(\bar{f}^{-i}, h_i)(t)) - \pi_i(s)$$
$$\leqslant m_j R_i - \left(\left[\frac{m_j \cdot R_i}{L_i}\right] + 1\right) L_i \leqslant 0.$$

Remark. Similarly we can prove that if $s^0,\ldots, s^{k-1} \in S$ satisfy $v_i < (1/k) \sum_{j=0}^{k-1} \pi_i(s^j)$ for all i, then there exists a strategy f which is a perfect equilibrium such that $\sigma(f)(t) = s^{t (\mathrm{mod}\ k)}$.

Let $s \in S$ be a strongly forced outcome such that for any player j who can profitably deviate from s (i.e. there exists $t_j \in S_j$ such that $\pi_j(s^{-j}, t^j) > \pi_j(s)$) there exists $w \in \operatorname{Conv} \pi(S)$ such that $w_i > v_i$ for all i and $\pi_j(s) > w_j$. Then there exists $f \in F$, a perfect stationary equilibrium, such that $\hat{\sigma}(f) = s$.

References

1. R. J. Aumann, Acceptable points in general cooperative n-person game, *in* "Contributions to the Theory of Games," (A. W. Tucker and R. C. Luce Eds.) Vol. IV, pp. 287–324, *Annals of Math. Studies* No. 40, Princeton Univ. Press, Princeton, N.J., 1959.
2. R. J. Aumann, Acceptable points in games of perfect information, *Pacific J. Math.* **10** (1960), 381–417.
3. R. J. Aumann, The core of a cooperative game without side payments, *Trans. Amer. Math. Soc.* **98** (1961), 539–552.
4. R. J. Aumann, A survey of cooperative games without side payments, *in* "Essays in Mathematical Economics in Honor of Oskar Morgenstern," (M. Shubick, Ed.), pp. 3–27, Princeton Univ. Press, Princeton, N.J., 1967.
5. R. J. Aumann, Lectures on game theory, Stanford University, 1976.
6. R. J. Aumann and L. Shapley, Long term competition—A game theoretic analysis, unpublished manuscript, 1976.
7. W. A. Brock, An axiomatic basis for the Ramsey Weizsacker overtaking criterion, *Econometrica* **38** (1970), 927–929.
8. J. W. Friedman, A non-cooperative equilibrium of supergames, *Internat. Econ. Rev.* **12** (1971), 1–12.
9. M. Kurz, Altruistic equilibrium, *in* "Economic Progress, Private Values, and Public Policy: Essays in Honor of William Fellner" (B. Balassa and R. Nelson, Eds.), pp. 177–200, North-Holland, Amsterdam, 1977.
10. M. Kurz, Altruism as an outcome of social interaction, *Amer. Econ. Rev.* **68** (1978), 216–222.
11. R. P. Luce and H. Raiffa, "Games and Decisions," Wiley, New York, 1957.
12. A. E. Roth, Self supporting equilibrium in the supergame, unpublished, 1975.
13. A. Rubinstein, "Equilibrium in Supergames," Research Memorandum No. 25, Center for Research in Math. Economics and Game Theory, The Hebrew University, Jerusalem, May 1977.
14. A. Rubinstein, Strong perfect equilibrium in supergames, unpublished, 1978.