Western University
Scholarship@Western

Centre for the Economic Analysis of Property Rights. Economics and Law Workshop Papers

Centre for the Economic Analysis of Property Rights

1982

The Most Durable Self-Enforcing Agreement

Lester G. Telser

Follow this and additional works at: https://ir.lib.uwo.ca/economicsceapr_el_wp

 Part of the Economics Commons

Citation of this paper:

Telser, Lester G.. "The Most Durable Self-Enforcing Agreement." Centre for the Economic Analysis of Property Rights. Economics and Law Workshop Papers, 82-16. London, ON: Department of Economics, University of Western Ontario (1982).

ARCC

ECONOMICS AND LAW WORKSHOP
82-16

THE MOST DURABLE SELF-ENFORCING AGREEMENT

Lester G. Telser

4:30 p.m.

Room 4161 SSC October 14, 1982

CENTRE FOR ECONOMIC ANALYSIS OF PROPERTY RIGHTS

ARCC
KF
801
.T44
1982

THE UNIVERSITY OF WESTERN ONTARIO

ECONOMICS AND LAW WORKSHOP
82-16

THE MOST DURABLE SELF-ENFORCING AGREEMENT

Lester G. Telser

4:30 p.m.

Room 4161 SSC October 14, 1982

Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637

July, 1981 Revised March, 1982 Preliminary

Please do not quote without the permission of the author.

Second Revision, September, 1982

Major funding for the Centre for Economic Analysis of Property Rights has been provided by the Academic Development Fund, The University of Western Ontario. Additional support has come from The Bureau of Policy Coordination, Consumer and Corporate Affairs. The views expressed by individuals associated with the Centre do not reflect official views of the Centre, The Bureau of Policy Coordination, or The University of Western Ontario.

Subscriptions to the Workshop papers and the Working Paper Series are $40 per year for institutions and $25 per year for individuals. Individual copies, if available, may be purchased for $3 each. Address all correspondence to John Palmer, Centre for Economic Analysis of Property Rights, The University of Western Ontario, London, Ontario, CANADA N6A 5C2.

833229040

KF
801
.T44
1982

The Most Durable Self-Enforcing Agreement

Lester G. Telser

1. Introduction

A self-enforcing agreement among n individuals continues only as long as each one believes faithful adherence to its terms is more profitable to him than a violation. Should such an agreement come to a halt owing to a violation by any one of the n parties to it, then each suffers only that loss given by the expected present value of their returns had the agreement gone on. During the time when an individual violates the agreement, the others suffer a loss that lowers their return below what it would have been in the absence of an agreement. At each point in time an individual can calculate his return from a current violation of the agreement and his loss which is the expected present value of his future returns under the agreement. Whenever his current return from a violation of the agreement exceeds the expected present value of his future returns from continuing the agreement, he will violate it. A violation of a self-enforcing agreement has only this kind of penalty. There is no intervention by third parties who can impose penalties on the violators.

An agreement among n individuals states the actions each must take. The return to an individual depends on the actions of all n individuals. An efficient outcome is a set of returns, one to each individual, that is a feasible result of their actions and is undominated by any other set of feasible results. Hence a set of returns is efficient if and only if there is no other set of returns that all of the n individuals would prefer. Usually there are many sets of efficient returns and among these are many candidates for a self-enforcing agreement. This fact, therefore, raises the question of which efficient outcome will the group choose. A principal result of this paper is to show when the efficient outcomes include a most-durable self-enforcing agreement and to derive some of its properties.

This theory distinguishes between induced and autonomous stopping. The former refers to stopping an agreement owing to a violation by one of the participants. The latter refers to a random event that changes the underlying circumstances enough so that it is no longer possible for the agreement to continue. Autonomous stopping is outside the control of the participants while the participants themselves can induce stopping as a means of punishing violations of a self-enforcing agreement.

A necessary feature of the theory of self-enforcing agreements is the uncertainty about the time of autonomous stopping. The theory assumes that the circumstances will continue for a finite number of periods and that the date when they stop is a random variable. Suppose, on the contrary, that the stopping date were known for sure to all of the participants. At the time of the last period cheating would be more profitable than faithfulness to the agreement. This is because there would then be no sacrifice of future returns to subtract from the current gains of a violation. Hence everyone would violate the agreement on the last date. But then the same argument applies to the next to the last date, the one before that and so on to the first date. We may conclude that there can be no self-enforcing agreement over a fixed number of time periods. This is the conclusion of Cournot in our current usage. With two or more firms producing a perfect substitute over a known finite horizon, collusion in order to secure the joint profit maximum would not occur according to his theory. However, in the presence of uncertainty about the duration of the demand for the product so that each firm is subject to the probability of future loss as punishment for present gain, it may well be that self-interest together with each individual's desire for maximum profit can result in collusion among them as a self-enforcing agreement.[1]

The origin of the theory of self-enforcing agreements is an attempt to explain the conditions under which a group of firms will find collusion more

profitable than competition to each of them individually even when they take into account the returns from cheating. The theory has a wider scope than this. It applies not only to a group of participants who have no direct transactions with each other and who affect each other only by their transactions with a common group of customers, but also to participants who deal directly with each other. For example, a buyer may buy repeatedly from a certain supplier only if he had satisfactory experience from the preceding transactions. Here a self-enforcing agreement is equivalent to an implicit long term contract that is subject to termination by either party without prior notice.

The theory given here extends my previous work on this subject (1972, 1978, and 1980) not only because it distinguishes a particular efficient outcome but also because it treats the problem for n instead of two individuals. This raises another important problem. Let an individual contemplate violation of the agreement. Can he do better with confederates than by acting alone? As we shall see, the answer is that he can do better acting alone. This result is useful since it implies that without loss of generality we can focus attention on a situation where a potential violator of an agreement calculates his expected gain that may result from cheating all of the other parties to it.

2. Self-Enforcing Agreements among n Individuals

Let u_i denote the return to individual i which depends on his own actions, y^i, and on the actions of the n - 1 other individuals in the group. The coordinates of y^i and z^i may represent the rates of sale of the individuals and u_i their net returns in applications where the n individuals are competing firms. Another important example is one in which the n individuals are workers who sell their labor services to a firm. In this application y^i refers to the quantity of labor services offered by worker i. The services are often complementary. As a third application we may think of a group of buyers and sellers so that u_i would represent the gain to individual i as a result of his transactions with some other individuals in the group. To express the relation between the return to an individual and his action and those of the others, write the function

$$u_i = f^i(y^i, z^i) \qquad i = 1, \ldots, n . \tag{1}$$

Each individual can choose the values of the coordinates of y^i independently of the choices made by the others of the coordinates of z^i that they control. Write the vector x to represent the actions of all n individuals as follows:

$$x = \langle y^1, y^2, \ldots, y^n \rangle$$

so that z^i is the complement of x with respect to y^i. More concisely, (1) becomes

$$u_i = f^i(x). \tag{2}$$

If each y^i has m coordinates then x has mn coordinates and is a point in a space of mn dimensions.

It is both reasonable and convenient to assume that each $f^i(x)$ is a concave and twice differentiable function of x. An outcome u is called efficient if it is feasible so that it is attainable by means of some action x and it is not dominated by any other feasible outcome. The latter means that for all feasible outcomes v, $v \not\geq u$ if u is efficient. Hence for at least one i,

$u_i > v_i$. Given the assumption of concave $f^i(x)$, an efficient outcome is a solution of the following maximum problem:

(3) Max $\Sigma\, \theta_i\, f^i(x)$ with respect to x for given $\theta_i \geq 0$ and $\Sigma\, \theta_i = 1$.

Assume this problem has at least one solution for some choices of $\theta = \langle \theta_1, \ldots, \theta_n \rangle$. This means the group has available more than one efficient outcome. Consequently, there is a potential conflict of interest among the members of the group because for any two distinct efficient outcomes u and v, $u \not> v$ and $v \not> u$.

Assume the n individuals agree to confine their choices of actions to those that give efficient results. Therefore an agreement determines θ and the x which depends on θ giving a solution of the maximum problem defined in (3). Under the agreement, individual i chooses the action y_o^i and the remaining individuals choose z_o^i, where $x_o^i = (y_o^i, z_o^i)$ denotes the optimal choice of x for the given θ so that x satisfies the necessary condition for a maximum,

(4) $$\Sigma\, \theta_i\, f_x^i(x_o) = 0,$$

where $f_x^i(.)$ denotes the partial derivative of $f^i(.)$ with respect to the coordinates of x. Under the terms of the agreement, the return to individual i would be

(5) $$u_{io} = f^i(y_o^i, z_o^i) .$$

If the n individuals do not reach an agreement because they are unable to do so or because they have broken a previous agreement, then each individual chooses his action in order to obtain the maximum return for himself given the choices of actions by all other individuals. This is to say the outcome is the noncooperative equilibrium x_N such that

(6) $$f^i(y^i, z_N^i) \leq f^i(y_N^i, z_N^i) \qquad i = 1, \ldots, n .$$

It is understood that each individual i chooses the best y^i for himself on his own. Concavity of $f^i(x)$ ensures the existence of at most one noncooperative equilibrium. For suitable additional conditions on $f^i(.)$, there is exactly one (Rosen, 1965).

A glance at Figure 1 will reveal the situation. The horizontal axis represents the return to the first individual denoted by the variable u and the vertical axis the return to all other individuals (or to a second individual) denoted by v. The curve LL' is the locus of efficient outcomes. The point E is a particular efficient point. The point N is the noncooperative equilibrium. Since it is inside the set of feasible outcomes and dominated by E, it is not an efficient point.

A violation of the agreement by individual i occurs if all other individuals remain loyal to it so that $z^i = z_o^i$ and individual i chooses his own action to maximize his own return given the actions of all the others. This is to say that individual i chooses y^i to solve the following problem:

(7) $\quad \text{Max } f^i(y^i, z_o^i) \quad$ with respect to y^i .

Notice that this is not the same as the result given in (6). In (6), the remaining n-1 individuals choose $z^i = z_N^i$ and individual i can do no better than to choose for his action y_N^i which is his component of the noncooperative equilibrium. The solution of the problem in (7) gives a different result. Individual i can obtain a higher return than $f^i(y_N^i, z_N^i) = u_N^i$ if the n - 1 other individuals choose an action consistent with the efficient point x_o while individual i chooses an action that is best for his own short-term self-interest. Figure 1 illustrates this. The coordinates of the point A give the return to the first individual if he cheats with respect to the efficient point E and the return to the others if they choose the actions to support E. Notice that at the point A the return to the cheating individual is above his return under the noncooperative equilibrium while the returns to his victims are below the level they would obtain under that equilibrium as well as being below the level under the efficient point E. It is also true that the cheating individual gets a higher return than he would under the efficient equilibrium.

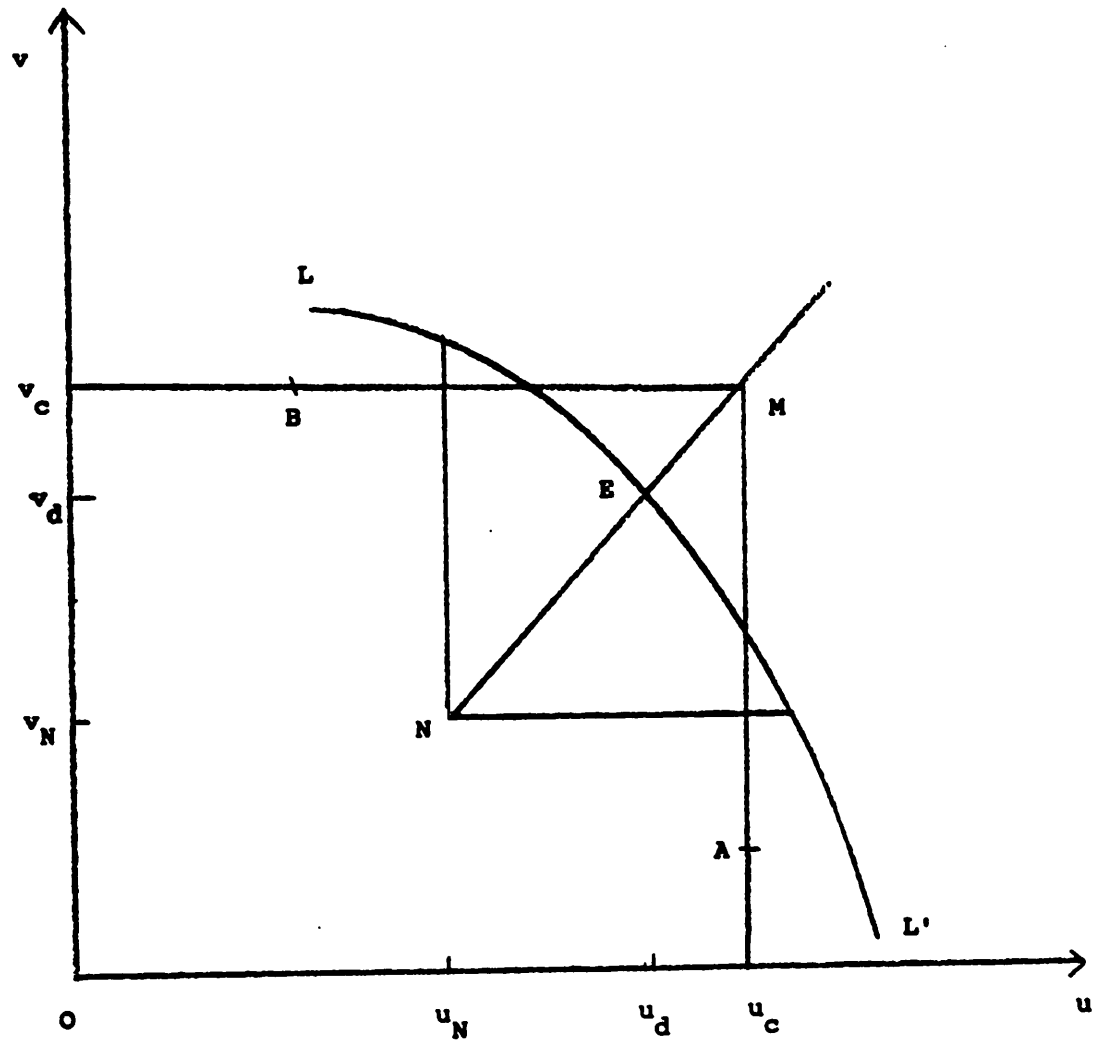

Figure 1

The efficient equilibrium E offers the individuals more than each would obtain under the noncooperative equilibrium N but it exposes them to the risk of a lower return than under N if one or more individuals cheat. It would seem from the figure that cheating is always more profitable to each individual than is adherence to E. The theory of self-enforcing agreements introduces new features in order to explain when a person best serves his interest by faithful adherence to E.

The new feature is the uncertainty of the duration of the underlying circumstances behind the common interest of the group. Think of these circumstances as if they result from a random process outside the control of the individuals. Let p_t denote the probability that these circumstances will last for t periods starting from t = o. Let T denote the random variable representing the duration of these circumstances. Hence

(8) Probability $(T = t) = p_t$ and Probability $(T > t) = q_t = \Sigma_{t+1}^{\infty} p_j$.

Therefore, stopping in period t has a probability of p_t and continuing for more than t periods has a probability q_t defined in (8). Stopping is certain eventually so that $\Sigma_{t=o}^{\infty} p_t = 1$. In particular, $p_o + q_o = 1$. Here p_o is the probability of never starting and q_o of going on for at least one period.

Next we need formulas to calculate the expected return when the duration of the underlying circumstances is a random variable. To simplify the notation, drop the subscript i from u_{it} and write

(9) $u_t = f(x_t)$

to denote the return to individual i in period t. If the process stops in period t, then $u_t = 0$ for all $t \geq t_o$ and for all x_t. Alternatively, we may assume that autonomous stopping lowers the return exogenously to some level not necessarily equal to zero. It is convenient for the sake of simplifying the notation to assume this level is zero. We refer to autonomous stopping, which is the outcome of a random event and not to that induced by a violation of the agreement.

Let s_t denote the sum of the returns up to and including period t .

(10) $$s_t = u_o + u_1 + \cdots + u_t .$$

The probability of receiving this sum is p_{t+1}. The expected return is the sum of $p_{t+1}\, s_t$ over all t from t = o to t = ∞. Call this expected return E(s) so that

(11) $$E(s) = \Sigma_o^{\infty} p_{t+1}\, s_t .$$

Replacing s_t with its components by means of (10) gives

(12) $$E(s) = u_o \Sigma_1^{\infty} p_j + u_1 \Sigma_2^{\infty} p_2 + u_2 \Sigma_3^{\infty} p_j +$$

(13) $$= u_o q_o + u_1 q_1 + u_2 q_2 + \cdots .$$

which uses (8) to go from (12) to (13). Understandably one may wish to call E(s) the expected return but it is not the expected value of the u_t's because q_t is not the probability of u_t. A necessary condition for a self-enforcing agreement is that p_t is strictly positive for all save possibly a finite number of periods. Also observe that since $\Sigma\, p_t = 1$ and $p_t \geq 0$, it follows that p_t approaches zero as t approaches infinity. It does not follow from this that q_t approaches zero as t approaches infinity.

We now wish to calculate the expected return of a party to a self-enforcing agreement. In doing so we can obtain a necessary and a sufficient condition for continuous loyalty to a self-enforcing agreement. The most costly penalty that a party to an agreement can impose on a violator is to choose the action giving the noncooperative equilibrium and refuse to resume the agreement after the discovery of a violation. If the gain of a violation exceeds the penalty under this arrangement, it surely exceeds the cost of less severe penalties. Therefore, study of the viability of a self-enforcing agreement under the most severe penalty gives a sufficient condition for its viability under less severe penalties. An example of a less severe penalty is one that suspends the agreement for a finite number of periods. Plainly such a penalty raises the expected gain of violations.

If individual i chooses $y^i \neq y_o^i$ so that there is a violation of the agreement, then, subsequently, each of the other individuals choose their actions noncooperatively. Hence following a violation of the agreement, $y^i = y_N^i$ for each i. Write

(14) $$\Sigma_o^{\infty} q_t \, f^i(y_o^i, z_o^i) ,$$

(15) $$\Sigma_o^{\infty} q_t \, f^i(y_N^i, z_N^i) ,$$

(16) $$\Sigma_o^{t-1} q_j \, f^i(y_o^i, z_o^i) + q_t \, f^i(y_c^i, z_o^i) + \Sigma_{t+1}^{\infty} q_j \, f^i(y_N^i, z_N^i) ,$$

where y_c^i denotes the optimal cheating action if x_o is the efficient action. Thus, y_c^i solves the maximum problem of (7). Expression (14) is the expected return of individual i if he is faithful to the cooperative agreement x_o in every period. The next expression (15) shows the expected return if there is no cooperation. Therefore, it gives the expected return of individual i under the noncooperative equilibrium. Expression (16) shows the expected return of individual i if he is faithful to the agreement from period o to period t-1, that is, for the first t periods, violates it in period t, and thereafter suffers the consequences of obtaining his return under the noncooperative equilibrium. This gives the lowest possible expected return to a violation of the agreement in period t because it assumes the violation lasts for only one period before discovery and there is relentless punishment in the form of noncooperation forever after. A less severe penalty would raise the expected return from violation. It would add no substantive issues to consider situations where violations last longer and are punished less severely (see Telser, 1972, chap 5).

The expected gain from cheating is the expression (16) minus the expression (14). It is more profitable to remain loyal to the agreement than to cheat if the expected gain from cheating is nonpositive for every t.

(17) $$q_t[f^i(y_c^i, z_o^i) - f^i(y_o^i, z_o^i)] - \Sigma_{t+1}^{\infty} q_j[f^i(y_o^i, z_o^i) - f^i(y_N^i, z_N^i)] \leq 0$$

is a sufficient condition for continuous loyalty to the cooperative equilibrium.[3]

This condition (17) is the principal focus of much of the subsequent analysis.

DEFINITION: The sequence $< q_t >$ supports a self-enforcing agreement for any efficient x_o for which (17) is true for all i and for all t.

Any efficient x that some $< q_t >$ can support is a candidate for a self-enforcing agreement. There is one x in particular that is a self-enforcing agreement for all possible $< q_t >$; this is x_N, the noncooperative equilibrium, which is usually inefficient. In general an efficient x is supportable as a self-enforcing agreement for some but not for all sequences $< q_t >$. As we shall see, if the expected duration of the circumstances underlying a self-enforcing agreement is long enough, then any efficient x can be a self-enforcing agreement. The plethora of candidates for these agreements explains our interest in special efficient points that are the most likely candidates such as the one which is the most durable.

We begin the analysis of (17) by simplifying it. Define μ_t as follows:

$$\mu_t = (\Sigma_{t+1}^{\infty} q_j)/q_t \quad . \tag{18}$$

In the most important special case, μ_t gives the expected duration of the circumstances underlying a self-enforcing agreement. Moreover, the expected duration is the same for all t. The special case is the one where the probability of autonomous stopping is the same for every period. In this case the probability of continuing for more than t periods is given by $(1 - p_o)^{t+1}$ where p_o = Probability [T = o] . Hence

$$q_t = (1 - p_o)^{t+1} \quad . \tag{19}$$

The random event bringing the underlying circumstances to a halt is independent of the preceding history. Think of tossing a coin where heads corresponds to the event, continuing, and tails to the event, stopping. The probability of a t-period horizon starting in the initial period o would correspond to a sequence of t consecutive heads followed by one tail. In this case μ_t does not

3. Properties of the Self-Enforcing Agreement among n Types of Individuals

We now consider whether two or more individuals together can each gain more by cheating the others than they would each gain by his own violation of the agreement. It is convenient to study this problem in a more general setting. This requires explicit treatment of coalitions among the n individuals. To this end, assume there are s_i individuals of type i instead of only one of each type as in the preceding section. Think of a large number of individuals of each type so that s_i is a real positive number, not necessarily an integer. The group of all individuals is represented by the n-vector S:

$$S = < s_1, s_2, \ldots, s_n > .$$

A coalition is some subset of the whole group that we may represent by an n-vector $R \leq S$ so that r_i with $0 \leq r_i \leq s_i$ represents the number of individuals of type i who belong to R. The complement of R includes everyone who is not a member of R. Denote the complement of R by CR so that

$$R + (CR) = S .$$

(Formally, we have a nonatomic game with n types of individuals.)

Let the coalition R contemplate a violation of the agreement by means of actions that would cheat the complementary coalition, CR. Coalition R has available the strategy y^R consisting of the Cartesian product of the strategies avilable to each of its members while the complementary coalition has available to it the strategies of its members, z^{CR}. Under the agreement R is supposed to choose y_o^R and CR to choose z_o^{CR}. There is no special affinity among the members of R; this coalition may be any subset of the whole group S, and the results apply to any R. For instance, it would make no difference whether the members of R are producers of complements or of substitutes.[5] If R cheats CR, it chooses y_c^R while if CR cheats R it chooses z_c^{CR}. Since u_i denotes the return to an individual of type i, the return to the coalition

R is $\Sigma\, r_i\, u_i$. There is the slight complication that the efficient action gives the maximum of $\Sigma\, \theta_i s_i u_i$ so we shall write $v_i = \theta_i u_i$ as an appropriate transformation of the returns to individuals of type i. This makes the return to the coalition R become $\Sigma\, r_i v_i$ (recalling that $r_i = 0$ means that individuals of type i do not belong to R). We now prove our first result.

LEMMA 1: Let $x_o = < y_o^R, z_o^{CR} >$ give the maximum of $\Sigma\, s_i v_i$, where $v_i = \theta_i u_i$ and $u_i = f^i(x)$. Let $< y_c^R , z_o^{CR} >$ give the maximum of $\Sigma_i\, r_i v_i$ with respect to y^R given $z^{CR} = z_o^{CR}$. Let $< y_o^R , z_c^{CR} >$ give the maximum of $\Sigma_i (s_i - r_i) v_i$ with respect to z^{CR} given $y^R = y_o^R$. It follows that

$$(1) \qquad \Sigma_i\, s_i f^i(y_o^R, z_o^{CR}) \leq \Sigma_i r_i f^i(y_c^R, z_o^{CR}) + \Sigma_i (s_i - r_i) f^i(y_o^R, z_c^{CR}) .$$

PROOF: By hypothesis, $x_o = (y_o^R, z_o^{CR})$ solves the following problem:

$$(2) \qquad \text{Max } \Sigma\, s_i f^i(y^R, z^{CR}) = \text{Max } [\Sigma\, r_i f^i(y^R, z^{CR}) + \Sigma\, (s_i - r_i) f^i(y^R, z^{CR}) .$$

with respect to y^R and z^{CR}. Also by hypothesis,

$$(3) \qquad \text{Max } \Sigma\, r_i f^i(y^R, z_o^{CR}) = \Sigma\, r_i f^i(y_c^R, z_o^{CR})$$

with respect to y^R and

$$(4) \qquad \text{Max } \Sigma\, (s_i - r_i) f^i(y_o^R, z^{CR}) = \Sigma\, (s_i - r_i) f^i(y_o^R, z_c^{CR})$$

with respect to z^{CR}. The feasible y for problem (3) include y_o^R and the feasible z for problem (4) include z_o^{CR}. Since (y_o^R, z_o^{CR}) is feasible for both problems (3) and (4), the solutions of these problems must satisfy

$$(5) \qquad \Sigma_i\, r_i f^i(y_c^R, z_o^{CR}) \geq \Sigma_i\, r_i f^i(y_o^R, z_o^{CR}) ,$$

and

$$(6) \qquad \Sigma_i (s_i - r_i)\, f^i(y_o^R, z_c^{CR}) \geq \Sigma_i (s_i - r_i) f^i(y_o^R, z_o^{CR}) .$$

Inequalities (5) and (6) in conjunction with (2) establish (1). QED.

An algebraic proof of this result is necessary because it is not evident from the geometry of Figure 1. The point A gives the outcome if R cheats RC and B the outcome if RC cheats R. The Lemma says that the sum of the vectors A and B gives a point northeast of the line tangent to the point E. It is also useful

to define the function W(R) so that we may see more clearly the meaning of condition (1).

(7) $$W(R) = \text{Max } \Sigma\, r_i\, f^i(y^R, z_o^{CR}) \quad \text{with respect to } y^R .$$

W(R) shows the maximum return to the coalition R for its optimal cheating strategy against the complementary coalition CR which is faithful to the cooperative agreement. Lemma 1 asserts that for all S such that S = R + CR,

(8) $$W(S) \leq W(R) + W(CR) ,$$

which means that W(.) is a subadditive function. We next prove that

LEMMA 2: W(R) is homogeneous of degree one.

PROOF: It suffices to show that

(9) $$W(R) = \Sigma\, r_i\, [\partial W(.)/\partial r_i] .$$

This is immediate because y_c^R is the optimal choice for R against z_o^{CR}, which gives the implication that

$$\partial W(R)/\partial r_i = f^i(y_c^R, z_o^{CR}) .$$

In conjunction with the definition of W(R), this gives the desired conclusion. QED.

With the aid of these results we can prove

THEOREM 1: W(R) is a convex function of R.

PROOF: Lemma 1 asserts W(R) is subadditive and Lemma 2 that it is homogeneous of degree one. Hence for all α such that $0 \leq \alpha \leq 1$,

$$W[\alpha R + (1-\alpha)P] \leq W(\alpha R) + W[(1-\alpha)P] = \alpha W(R) + (1-\alpha)W(P) . \quad \text{QED}$$

By virtue of the convexity of W(R), we may conclude that a coalition wishing to obtain the maximal return from cheating the others will include all the individuals of a given type. Nor is this all. Some particular type can gain the most by cheating the others. It follows that cheating is least likely to occur among similar types of individuals. Since it never pays for different types of individuals to join together in order to cheat the others owing to convexity, a potential violator of the agreement gains the most when he acts alone.

For this reason, study of the viability of a self-enforcing agreement may confine its attention to the possible gains of a single individual who calculates whether loyalty to the agreement best serves his interest. This result is plausible. A potential violator acting alone has the most victims to plunder.

The next result asserts that any efficient strategy, say x_o, is supportable as a self-enforcing agreement if the expected horizon, μ, is long enough. Reconsider (2.22) and define the expected gain from violating the agreement by the function $F^i(.)$ as follows:

(10) $$F^i(.) = f^i(y_c^i, z_o^i) - f^i(y_o^i, z_o^i) - \mu_i[f^i(y_o^i, z_o^i) - f^i(y_N^i, z_N^i)].$$

If $F^i(.) > 0$, individual i gains more from cheating than from loyalty. He gains more from loyalty than cheating when $F^i(.) \leq 0$. Therefore, the sign of $F^i(.)$ determines whether or not it is in the self-interest of individual i to abide by the agreement. Agreement among all of the n individuals is optimal provided $F^i(.) \leq 0$ for each i. We now have a more general case since μ_i has a subscript. This means the expected horizon that ensures loyalty is not necessarily the same for all individuals. The critical expected horizon is the largest μ_i because loyalty is more profitable than disloyalty if

$$E(T) \geq \operatorname*{Max}_{\mu_i} \{ \mu_i : F^i(.) = 0 \}.$$

This leads us to study the properties of $F^i(.)$ in order to learn how disparities among the individuals affect the feasibility of a self-enforcing agreement among them. Rewrite (10) in the following more convenient form:

(11) $$F^i(.) = f^i(y_c^i, z_o) - (1 + \mu_i) f^i(y_o, z_o) + \mu_i f^i(y_N, z_N),$$

dropping the superscript i from the action variables, which should cause no ambiguity. Recall that each $f^i(.)$ is a concave function. This implies

$$[1/(1+\mu_i)]f^i(y_c, z_o) + [\mu_i/(1+\mu_i)] f^i(y_N, z_N) \leq f^i(y^*, z^*),$$

where

$$y^* = [1/(1+\mu_i)] y_c + [\mu_i/(1+\mu_i)] y_N ,$$

$$z^* = [1/(1+\mu_i)]\, z_o + [\mu_i/(1+\mu_i)]\, z_N \ .$$

Consequently,

(11) $$F^i(.) \leq (1+\mu_i)\, [f^i(y^*, z^*) - f^i(y_o, z_o)] \ .$$

By hypothesis, $x_o = (y_o, z_o)$ is feasible since it is the efficient choice for the group. The noncooperative equilibrium is also feasible, though inefficient. Since the set of feasible actions is convex, z^*, a weighted average of feasible actions,is itself feasible. Now y^* is feasible provided μ_i is large enough because if this is true then y^* is close enough to the feasible y_N . Hence the right side of (11) is nonpositive for large enough μ_i . Choose μ equal to the largest of the μ_i's that are each large enough to make the right side of (11) nonpositive for each i. Consequently,

(12) $$\Sigma\, s_i F^i(.) \leq (1 + \mu)\, \Sigma\, s_i [f^i(y^*,z^*) - f^i(y_o,z_o)] \leq 0$$

where $\mu = \text{Max} < \mu_i >$. This completes the proof of

THEOREM 2: For all large enough μ, x_o is a supportable self-enforcing agreement

We can now show how similarity among the n types of individuals is conducive to a self-enforcing agreement. To this end, write

(13) $$f^i(y^i, z^i) = f(y^i, z^i, a^i) + \delta,$$

where a^i is a vector with a finite number of coordinates chosen so that the common function f(.) gives an approximation to the individual functions $f^i(.)$ with an error δ. Given that each function $f^i(.)$ is continuous, the Weierstrass Theorem allows us to approximate uniformly on a suitable closed domain the n functions $f^i(.)$ by the same continuous function f(.). Differences among the individual types are expressed as coordinates of the vector a^i. This induces a similar approximation for the functions $F^i(.)$ which represent the expected gains from cheating. We have

(14) $$F^i(.) \approx F(.\ , a^i) \ .$$

We can now prove

COROLLARY: Define Δ so that $\Delta \geq \|a^i - a\|$. If Δ is small enough and the expected horizon μ is long enough then $F^i(.) \leq 0$ for each i so that cheating is less profitable than loyalty for each individual type.

PROOF: If Δ is small enough, then each $F(\ .\ ,\ a^i)$ is close to $F(.\ ,\ a)$. By Theorem 2, for μ large enough, $F(.\ ,\ a) \leq 0$. Therefore, for each a^i $F(.\ ,\ a^i) \approx F(.\ ,\ a) \leq 0$ and since $F^i(.) \approx F(.\ ,\ a^i) \leq 0$, this gives the conclusion. QED

This result confirms our intuition that if all members of a group are sufficiently alike, then none will gain by cheating if the expected horizon is long enough. Similarity among the individuals alone is not sufficient to give a self-enforcing agreement. It must also be true that the expected horizon is long enough.

Theorem 2 gives an embarassment of riches. It asserts that any efficient action giving each type of individual a return better than he would get under the noncooperative equilibrium is a possible self-enforcing agreement for a long enough expected horizon. But not all efficient actions are equally plausible candidates for a self-enforcing agreement. This leads to the problem of how to strengthen the theory in order to select a particular efficient action as the most likely choice of the group.

As we have seen, the noncooperative equilibrium is always self-enforcing no matter how short is the expected horizon but it is usually inefficient. There are inefficient points that dominate the noncooperative equilibrium and everyone would be better off with these than with their returns under the noncooperative equilibrium. Nor is this all. These inefficient points that dominate the noncooperative equilibrium are self-enforcing for long enough expected horizons. The closer an inefficient point is to the locus of efficient points, the more durable must be the underlying circumstances that will make that point self-enforcing. Efficient points as candidates for a self-enforcing

agreement require the longest expected horizon. This fact constitutes one of the most persuasive arguments in favor of choosing that efficient point with the longest expected horizon, namely, the most durable efficient point as the best candidate for a self-enforcing agreement.

4. The Most Durable Self-Enforcing Agreement

Thanks to Theorem 1, we can simplify the analysis of the gains of cheating by considering the case of two individuals. The first individual, call him the f-player, has a return given by the function

(1) $u = f(y, z)$,

where y denotes his action and z the action of the g-player whose return is given by the function as follows:

(2) $v = g(y, z)$.

As in the preceding sections we shall assume both functions are concave, twice differentiable and have a unique noncooperative equilibrium given by the (inefficient) pair (y_N, z_N) . There are efficient pairs, (y, z), giving the maximum of $\theta u + (1-\theta)v$ for an agreed-upon choice of θ in a range contained in the unit interval. In fact, the most-durable self-enforcing agreement determines θ and this induces the corresponding returns to the players. The following functions give the expected gains from cheating to the f and g players:

(3) $F(.) = f(y_c, z) - f(y, z) - \lambda [f(y, z) - f(y_N, z_N)]$,

(4) $G(.) = g(y, z_c) - g(y, z) - \mu [g(y, z) - g(y_N, z_N)]$,

where (y, z) denotes the efficient pair of actions corresponding to the agreed-upon θ . The optimal cheating action for the f and g players satisfy

(5) $f_y(y_c, z) = 0$ and $g_z(y, z_c) = 0$.

The efficient pair (y, z) satisfies

(6) $\theta f_y(y, z) + (1 - \theta) g_y(y, z) = 0$,

(7) $\theta f_z(y, z) + (1 - \theta) g_z(y, z) = 0$.

It follows from (5) that y_c depends on z and z_c depends on y. This means that y_c, the optimal action for the f-player who wishes to cheat the g-player, depends on the efficient action of the g-player. However, the noncooperative pair of actions (y_N, z_N) does not depend on the efficient actions. Hence for different values of θ, there are different efficient actions (y, z) and

these in turn determine y_c and z_c. It is an implication of Lemma 1 that

(8) $$f(y_c, z) \geq f(y, z) \quad \text{and} \quad g(y, z_c) \geq g(y, z) .$$

The parameters λ and μ represent the expected horizons for the f and g players respectively. For all values of λ such that $F(.) \leq 0$ the f-player expects a higher return from loyalty than from cheating. Similarly, for all values of μ such that $G(.) \leq 0$, loyalty promises a higher return to the g-player than does cheating. The smallest values of these parameters that are each large enough to make loyalty the more profitable course are those for which $F(.) = 0$ and $G(.) = 0$. Theorem 2 guarantees the existence of λ and μ that can satisfy these equations for a prescribed efficient pair of actions and the induced optimal cheating actions. A self-enforcing agreement is feasible if and only if the expected horizon exceeds the larger of the two values λ and μ for which $F(.) = 0$ and $G(.) = 0$.

The nature of the argument is easier to grasp with the aid of Figure 2. The vertical axis shows λ and μ while the horizontal axis shows $u - u_N$ which is the excess of the f - player's return under an efficient point over his return under the noncooperative equilibrium. The curve AB gives the locus of pairs $(\lambda , u - u_N)$ for which $F(.) = 0$. Now in place of (3) write

(8) $$\lambda = (u_c - u)/(u - u_N) .$$

where $u_c = f(y_c, z)$. The λ defined in (8) makes $F(.) = 0$. The larger is $u - u_N$, the smaller is $u_c - u$. Hence the larger is $u - u_N$, the smaller is the ratio on the right side of (8). This means the expected horizon that is necessary and sufficient for $F(.) = 0$ varies inversely with $u - u_N$. The points above the curve AB have $F(.) < 0$ and those below AB have $F(.) > 0$. Hence only expected horizons that lie above the curve AB can support a self-enforcing agreement for the f-player. For a given expected horizon, say $E(T)_1$, any excess of $u - u_N$ lying to the right of the point C is an acceptable self-enforcing agreement to the f-player.

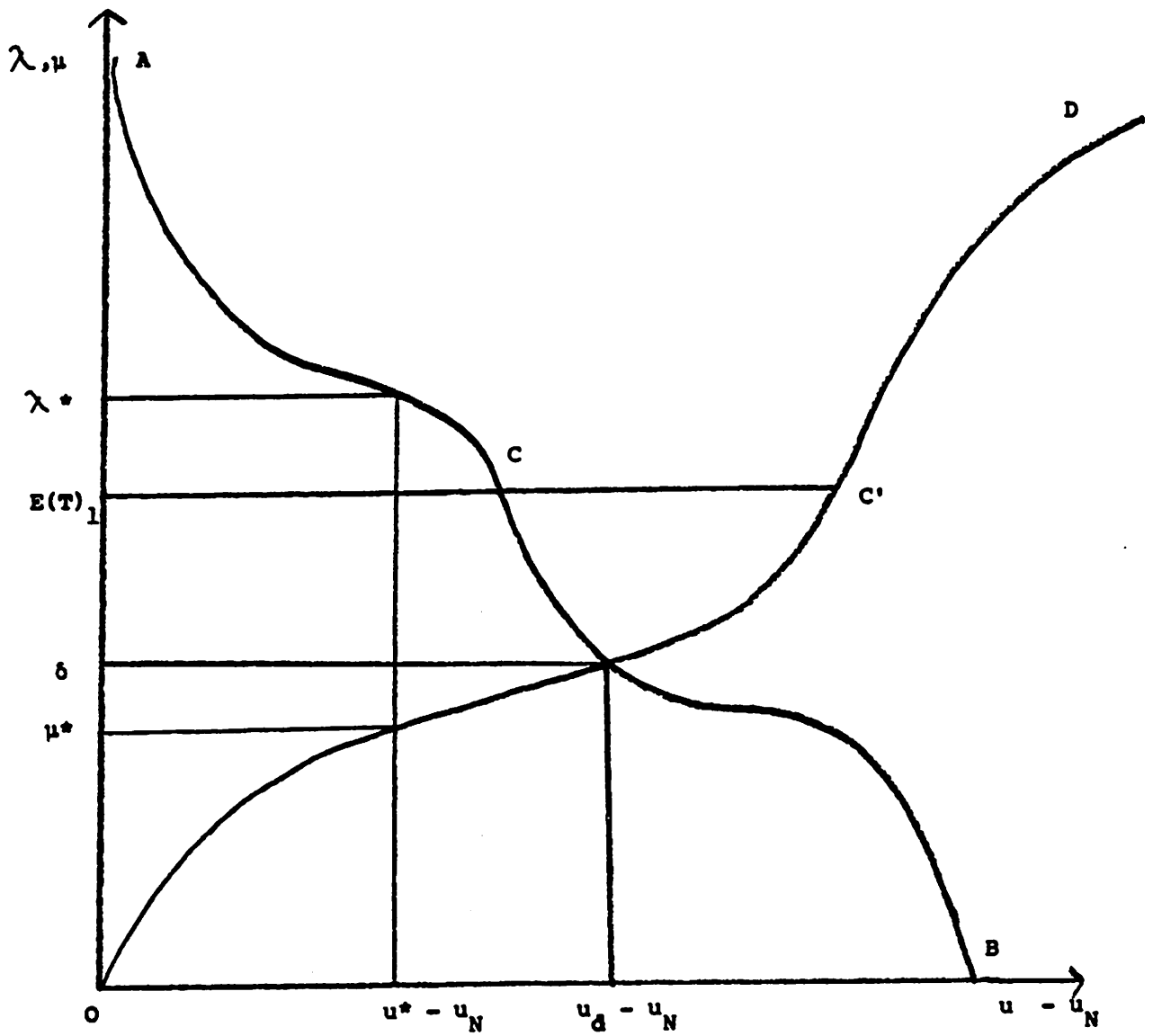

Figure 2

Figure 2 also shows the nature of the relation between μ and $u - u_N$ that makes $G(.) = 0$ for the g-player. From (4) we obtain

(9) $$\mu = (v_c - v)/(v - v_N) \qquad (v_c = g(y, z_c))$$

that gives the expected horizon for the g-player that is necessary and sufficient to make cheating and loyalty equally profitable. Along the efficient locus of returns, u and v vary inversely (see fig. 1). Consequently, since (u_N, v_N) is a given point, $v - v_N$ also varies inversely with $u - u_N$. We may conclude that μ varies directly with $u - u_N$. Therefore, the larger is $u - u_N$, the smaller is $v - v_N$ and the larger is $v_c - v$. Hence the larger is $u - u_N$, the larger must be the μ that is necessary and sufficient to sustain a self-enforcing agreement for the g-player.

In order to have a self-enforcing agreement between the f and the g players, the expected horizon must be greater than or equal to the larger of the two values λ and μ that makes the two of them indifferent between loyalty and disloyalty to a given agreed-upon efficient pair (u, v). Thus, in Figure 2, the expected horizon must exceed λ^* in order to make $u^* - u_N$ (and the corresponding $v^* - v_N$) the subject of a self-enforcing agreement. The shortest possible expected horizon that can support a self enforcing agreement decreases as $u^* - u_N$ moves to the right toward $u_d - u_N$. The expected horizon corresponding to the efficient return $u_d - u_N$ is the intersection of the two curves AB and OD; it is the most durable self-enforcing agreement because it can support the point $u_d - u_N$ for all expected horizons more durable than δ.

Figure 1 illustrates the relations between the noncooperative equilibrium, N, whose coordinates are (u_N, v_N), the most durable self-enforcing agreement, E with coordinates (u_d, v_d) and the two cheating points A, (u_c, v_d) and B, (u_d, v_c). Since the most durable self-enforcing agreement equates λ and μ as given by (8) and (9), the point E is on the line joining N to M, which has coordinates given by (u_c, v_c) as Figure 1 shows. M is not itself feasible.

Figure 1 makes plain certain properties of the most durable self-enforcing agreement. First, assume the players are alike. This means the locus of efficient outcomes is symmetric with respect to the 45°-line. In this case the three points, N, E and M all lie on the 45°-line so that the players' most durable self-enforcing agreement gives them each the same return. Second, the diagram shows that the player who can gain the most from cheating must obtain the largest return order to make the self-enforcing agreement most durable.[6]

5. Relation between the Most Durable Self-Enforcing Agreement and the Nash Bargaining Point

The Nash Bargaining Point, call it the NBP, is the solution of the following maximum problem:

Max $(u - u_N)(v - v_N)$ with respect to efficient pairs (u, v)

(Nash, 1950a, 1953). Call the solution (u_1, v_1). Call the pair for the most durable self-enforcing agreement, δ-SEA. We prove

THEOREM 3: $(u_c - u_d)(v_c - v_d) \leq (u_c - u_1)(v_c - v_1)$

so that the geometric mean of the expected gain from cheating under the δ-SEA does not exceed the geometric mean of the gain from cheating under the NBP.

PROOF: Since (u_d, v_d) gives the minimum of δ, it also gives the minimum of δ^2.

(1) $\delta^2 = [(u_c-u_d)/(u_d-u_N)][(v_c-v_d)/(v_d-v_N)] \leq [(u_c-u_1)/(u_1-u_N)][(v_c-v_1)/(v_1-v_N)]$.

The NBP satisfies the following inequality:

$$(u_1 - u_N)(v_1 - v_N) \geq (u_d - u_N)(v_d - v_N) ,$$

recalling the (u_d, v_d) is efficient and therefore feasible. Consequently,

(2) $[(u_c-u_1)/(u_1-u_N)][(v_c-v_1)/(v_1-v_N)] \leq [(u_c-u_1)/(u_d-u_N)][(v_c-v_1)/(v_d-v_N)]$.

Together (1) and (2) yield the desired conclusion. QED.

Like the NBP, the δ-SEA is invariant under linear transformations of the utility indicators so that von-Neumann-Morgenstern utility can measure the returns to the individuals. So we prove

THEOREM 4: The δ-SEA is invariant under linear tranformations of the utility indicators.

PROOF: Let u and v denote the original utility indicators and U, V the linear transforms so that

$$U = a_o + a_1 u , \qquad a_1 > 0 ,$$

$$V = b_o + b_1 v , \qquad b_1 > 0 .$$

Choose a_o, b_o so that U_N, V_N corresponds to u_N, v_N . Hence

$$U - U_N = a_1(u - u_N) \quad \text{and} \quad V - V_N = b_1 (v - v_N) .$$

Choose a_1, b_1 so that U_d, V_d corresponds to u_d, v_d. Consequently,

$$a_1 = (U_d - U_N)/(u_d - u_N) \quad \text{and } b_1 = (V_d - V_N)/(v_d - v_N) .$$

We now have

$$(3) \qquad (U - U_N)/(U_d - U_N) = (u - u_N)/(u_d - u_N)$$

and

$$(4) \qquad (V - V_N)/(V_d - V_N) = (v - v_N)/(v_d - v_N).$$

Since (u_d, v_d) is a δ-SEA, it follows from (4.8) and (4.9) that

$$(5) \qquad (u_c - u_d)/(u_d - u_N) = \delta = (v_c - v_d)/(v_d - v_N) .$$

We must show that (5) also holds for U, V so that

$$(6) \qquad (U_c - U_d)/(U_d - U_N) = \delta = (V_c - V_d)/(V_d - V_N) .$$

It follows from (5) that

$$(u_c - u_N + u_N - u_d)/(u_d - u_N) = \delta$$

which implies that

$$(u_c - u_N)/(u_d - u_N) = 1 + \delta = (U_c - U_N)/(U_d - U_N) ,$$

and

$$(U_c - U_N)/(U_d - U_N) - 1 = \delta .$$

Therefore, as desired,

$$(U_c - U_d)/(U_d - U_N) = \delta ,$$

and a similar result holds for V . QED.

The δ-SEA does not satisfy the Axiom of the Independence of Irrelevant Alternatives while the NBP does satisfy this axiom. (For a discussion of this axiom and its relation to the NBP, see Luce and Raiffa, 1957, sec. 6.5.)

THEOREM 5: The δ-SEA does not satisfy the Axiom of Independence with respect to Irrelevant Alternatives (I.I.A. Axiom).

PROOF: Plainly, the δ-SEA depends on u_c and v_c (and conversely). Therefore, any change in the set of feasible alternatives that removes u_c or v_c must also

change the δ-SEA. This is to say that the δ-SEA cannot satisfy the I.I.A. Axiom. QED.

Before explaining why this result is plausible, one should recognize that although the δ-SEA violates the IIA Axiom, it may nevertheless give the same outcome as the Nash Bargaining Point. For instance, this is true in the important special case in which all of the individuals are alike so that under the Nash Bargaining Point as well as the most durable self-enforcing agreement all would obtain the same return. With symmetry, therefore, the Nash Bargaining Point is also the most durable self-enforcing agreement.

The argument leading to the Nash Bargaining Point assumes none of the players violate the agreement. In the absence of an agreement there is the noncooperative equilibrium. Indeed no violations of the agreement can occur in the Nash bargaining theory because there is only one time period. In contrast, the theory of self-enforcing agreements makes essential use of a sequence of repetitions over a finite horizon of uncertain duration. The probability of punishment occurring after a violation of the agreement is crucial. The terms of the agreement depend not only onthe returns under the noncooperative equilibrium, true also of the Nash bargaining theory, but also on the temporary gains from violating the agreement, absent from the Nash bargaining theory. Therefore, while shrinking the set of feasible returns has no effect as long as the previously agreed-upon choice remains feasible according to the Nash bargaining theory, it does affect the outcome in the theory of the most durable self-enforcing agreement if it removes u_c or v_c or both.

6. Uncertain Finite Horizon or Infinite Horizon?

One may interpret E(s) defined in (2.13) as the present value of an income stream over an infinite number of periods with the t-period return, u_t, discounted by the factor q_t. On this interpretation a low discount rate corresponds to a long expected horizon. Instead of calling

(1) $$q_t(u_t^c - u_t^e) + \Sigma_{t+1}^{\infty} q_h(u_h^N - u_h^e) \quad ,$$

the expected gain from cheating, we call this the present value of the gain from cheating. Here u_t^c gives the return from cheating against the efficient point u_t^e. The two interpretations are formally equivalent. There is the question of whether they are operationally equivalent. The answer is in the negative.

Interest rates are higher when there is prosperity than when there is a lack of it. Therefore, according to the present-value-over-an-infinite-horizon interpretation, the incentive to violate a self-enforcing agreement would be greater when interest rates are high than when they are low. This is because the present value of the penalty of a violation is lower relative to the return from the violation when interest rates are high. Violations are more likely to occur during periods of high than of low interest rates. It follows, therefore, that a cartel agreement is more likely to break down during prosperous times than during recessions according to this theory. Moreover, cartels would be more likely to form during periods of recession and depression.

Since the risk of business failure is greater during periods of recession than during periods of prosperity, the expected duration of the circumstances resulting in a self-enforcing agreement is greater during periods of prosperity than recession. Hence under the interpretation that E(s) gives the expected return over finite horizons of uncertain duration, it would follow that cartels

are more likely to collapse during depressed times than during prosperous times. Also, cartels are more likely to form during prosperous than during unprosperous times. Hence the two theories seem to make opposite predictions about the liklihood of a self-enforcing agreement among a group of firms depending on whther there is prosperity or not.

These predictions seem vulnerable to the criticism that they ignore the effect of prosperity and depression on the returns themselves. But notice that the incentive to cheat depends on the difference between the returns from cheating, the penalty it invokes, and the return to loyalty. Granting this, such considerations would not distinguish between the two theories. This is because the pattern of differences $u_t^c - u_t^e$ and $u_h^N - u_h^e$ is the same for both theories. The probability of autonomous stopping is higher during depressed than during prosperous periods so that the q's are different in a way that distinguishes them from the cyclical pattern of interest rates.

7. Appendix

This section presents and proves several propositions that are important for the results in the text. First, we show that the noncooperative equilibrium is self-enforcing for all sequences $< q_t >$. Write the expected gain from cheating in period t as follows:

(1) $$F(.) = q_t\ (u_t^c - u_t^e) + \Sigma_{t+1}^{\infty}\ q_h(u_h^N - u_h^e)\ .$$

Here u^c denotes the return if the player cheats when the agreed-upon outcome is u^e and u^N denotes the noncooperative equilibrium. Cheating is less profitable than loyalty if and only if $F(.) \leq 0$. Plainly if the players do not cooperate and choose the noncooperative equilibrium so that their return is u^N in each period, then the term $u_h^N - u_h^e = 0$ for each period h and $u_t^c < u_t^e = u_t^N$. Hence $F(.) < 0$ if $u_t^e = u_t^N$ for every t.

The text asserts in several places that $F^i(.) \leq 0$ is necessary for a self-enforcing agreement. This deserves a proof. Assume there is a self-enforcing agreement. This means that loyalty is more profitable than disloyalty because the expected return from disloyalty is less than the expected penalty. If the penalty for disloyalty is the most severe so that disloyalty is punished by reverting to the noncooperative equilibrium even after a single violation then the desired conclusion is true. If on the other hand there is a less severe penalty for disloyalty, this deters cheating by virtue of the hypothesis there is a self-enforcing agreement. But then the most severe penalty a fortiori would also deter disloyalty. So in either case, if there is a self-enforcing agreement then $F^i(.) \leq 0$. This completes the proof of

THEOREM 6: $F^i(.) \leq 0$ is necessary for a self-enforcing agreement.

The third topic that requires our attention is necessary and sufficient conditions for the existence of the most-durable self-enforcing agreement. We seek conditions that ensure the existence of an intersection of the two curves in Figure 2.

The pairs of actions (y_c, z) and (y, z_c) are both feasible but not efficient. There is an efficient pair, call it (y_f, z_f) that is most favorable to the f-player and, consequently, least favorable to the g-player. Likewise, there is an efficient pair (y_g, z_g) most favorable to the g-player and least favorable to the f-player. For the pair (y_f, z_f), $(u_c - u_N)/(u - u_N)$ is a minimum and $(v_c - v_N)/(v - v_N)$ a maximum. Thus, the outcome most favorable to the f-player and least favorable to the g-player gives the latter the most incentive to cheat and the former the least incentive to cheat. A parallel argument applies for the pair (y_g, z_g).

Consider the function H(y, z) defined as follows:

(1) $$H(.) = (u_c - u_N)/(u - u_N) - (v_c - v_N)/(v - v_N) .$$

A most-durable self-enforcing agreement exists if and only if there is an efficient pair (y, z) for which H(y, z) = 0. The function H(.) measures the difference between the incentive to cheat by the f and the g players. In terms of Figure 2, it measures the vertical difference between the two curves AB and OD. This difference is a maximum where the f-player has the most incentive to cheat and the g-player the least incentive to cheat. It has a minimum where the reverse is true. The function H(y, z) is the difference between two monotonic functions that move in opposite directions on the domain of efficient pairs of strategies. The maximum of H(.) is at (y_g, z_g) and the minimum at (y_f, z_f). This motivates the following

DEFINITION: There is said to be weak parity between the f- and the g-players if Min H(y, z) with respect to efficient pairs (y, z) has the same sign as Min [-H(y, z)] over the same domain.

When there is weak parity, outcomes most favorable to either player are equally unstable because whichever player is the most favored, the other player has an incentive of the same sign to violate the agreement. The adjective "weak"

is appropriate because the incentive to cheat refers only to the sign and not to the magnitude.

It is easy to verify that there is weak parity between the two players if and only if Max H(y, z) and Max [-H(y,z)] have the same sign. Since

$$\text{Min}[-H(y,z)] = -\text{Max } H(y,z),$$

it follows that Min H(.) and Max H(.) are of opposite sign when there is weak parity. The main result is given in the following

THEOREM 7: Let H(y,z) be a continuous function of (y, z) on the domain of efficient pairs. Let there be weak parity between the f and the g players. Let Max H(y, z) > 0 on the domain of efficient pairs.

It follows there exists a most-durable self-enforcing agreement.

PROOF: Since there is weak parity and Max H(y, z) > 0, Min H(y, z) < 0. Continuity of H(.) therefore implies the desired conclusion. QED.

Reconsider the definition of H(y, z) in (1). Continuity of H(.) means that neither player can receive a return equal to what it would be in the noncooperative equilibrium. Efficiency precludes them both from getting the same return as they would under the noncooperative equilibrium. We may conclude that if (u_d, v_d) is the most-durable self-enforcing agreement then

$$u_d - u_N > 0 \quad \text{and} \quad v_d - v_N > 0.$$

There is a straightforward extension of the theory of the most durable self-enforcing agreement for n individuals. Such an agreement exists if there is an efficient $x = (y^i, z^i)$ such that for all $i = 1, \ldots, n$,

(2) $$(u_{ic} - u_i)/(u_i - u_{iN}) = \delta$$

Equivalently write

(3) $$(u_{ic} - u_{iN})/(u_i - u_{iN}) = 1 + \delta .$$

To study the conditions under which the n functions in (3) have a solution, introduce the n(n-1)/2 functions $H_{ij}(.)$ as follows:

(4) $$H_{ij}(.) = (u_{ic} - u_{iN})/(u_i - u_{iN}) - (u_{jc} - u_{jN})/(u_j - u_{jN}) .$$

The definition of weak parity applies to each pair of individuals. A most durable self-enforcing agreement exists if the hypotheses of Theorem 7 apply to each function $H_{ij}(.)$. From this we see once more that disparities among the n individuals are not conducive to the existence of a most durable self-enforcing agreement.

Footnotes

Roy Radner posed to me the question of which efficient outcome is the most likely candidate for a self-enforcing agreement. I am grateful to Gary S. Becker and George J. Stigler for their helpful comments.

1. Luce and Raiffa deserve credit for being the first to point out that an uncertain finite horizon can resolve the Prisoners' Dilemma (1957, p. 102). Both J. Friedman (1971 and 1977) and I (1972, 1978, and 1980) acknowledge our debt to them. Smale (1980) is an interesting recent contribution to this Dilemma. Radner (1980) gives a theory of the ε-equilibrium that attempts to solve the Prisoners' Dilemma by assuming they act as if the return is an average of the returns over a finite horizon. This has the strange property that the past returns are added together with the future returns in order to calculate the average. It is as if the individuals can lose their past returns if they violate the agreement in the present. It seems possible to make sense of this only if the individuals post a bond representing their accumulated past returns which they can lose if they violate the agreement. Since a third party would have to hold the bond and determine whether there has been a violation, Radner's device does not appear to be a self-enforcing agreement. His second definition of the return is an average of the profits over all of the remaining periods. However, in this case, cooperation breaks down as the last period approaches. Hence it does not resolve the problem of the Prisoners' Dilemma. For an explicit treatment of enforcing an agreement over a finite period of time by means of posting a bond see Becker and Stigler (1974). A recent contribution to self-enforcing agreements giving many applications is Klein and Leffler (1981).

2. The noncooperative equilibrium is a concept due to Nash (1950 and (1951). Hence I use the subscript N to denote the Nash point. Cournot (1838) introduced an important special case in his analysis of oligopoly.

3. The Appendix shows why this inequality is also necessary for a self-enforcing agreement. There is no loss of generality in conducting the analysis on the hypothesis of the most severe punishment following a violation of the agreement.

4. See the Appendix for a proof that the noncooperative equilibrium is self-enforcing for all sequences $< q_t >$.

5. We assume that the individuals can make independent choices of their actions. This precludes fixed proportions among their actions.

6. There are some formal similarities between my theory of the most durable self-enforcing agreement and J. Friedman's theory of balanced temptation of which I was unaware until after I had worked out my own ideas. Friedman's condition (8.9) (1977, chap. 8) corresponds to my condition (2.22). Instead of my expected horizon μ_i, he has $\alpha_i/(1 - \alpha_i)$, where α_i is the discount factor for player i so that $\alpha_i = 1/(1 + \rho_i)$, where ρ_i is the discount rate. Hence $\mu_i = 1/\rho_i$. Friedman calls the following ratio:

$$[f(y_c, z) - f(y, z)]/[f(y, z) - f(y_N, z_N)] ,$$

the temptation to cheat because it gives "the ratio of the one-period gain to the per period later loss" p. 180. He proposes the efficient equilibrium giving each player the same temptation to cheat. Hence there is an upper bound on the discount rate, call it ρ^*, such that if $\rho_i < \rho^*$, it will not pay for any player to cheat. He also points out that his equilibrium is invariant with respect to linear transformations of the utility function and that it does not satisfy the Axiom of the Independence of Irrelevant Alternatives. His earlier publication (1971) presents these ideas more abstractly and concisely.

His theory requires an infinite horizon in order to be valid. In contrast, my theory assumes a finite horizon of uncertain duration. It lays the stress on uncertainty. As the analysis in sec. 6 points out, the two theories make different predictions and are empirically distinguishable. The reasoning that

leads to the most durable self-enforcing agreement uses a different approach than that leading to a balanced temptation. Although I would want the level of interest rates to affect the terms of the most-durable self-enforcing agreement, it is also desirable to introduce explicitly the effects of uncertainty about the duration of the underlying circumstance. This furnishes a broader range of empirical insights.

REFERENCES

Becker, Gary S. and George J. Stigler. 1974. Law Enforcement, Malfeasance and Compensation of Enforcers. Journ. Legal Studies 3:1 - 18 (Jan).

Cournot, Augustin. 1960 Researches into the Mathematical Principles of the Theory of Wealth, trans. Nathaniel Bacon from 1838 French ed. New York: Kelley.

Feller, William. 1962. An Introduction to Probability Theory and Its Applications. 2d ed. vol. 1. New York: Wiley.

Friedman, James W. 1971. A Non-cooperative Equilibrium for Supergames. Rev. Econ. Studies 38:1-12. (Jan).

_____. 1977. Oligopoly and the Theory of Games. Amsterdam: North-Holland.

Klein, Benjamin and Keith B. Leffler. 1981. The Role of Market Forces in Assuring Contractual Performance. Journ. Pol. Econ. 89:615 - 41 (Aug).

Luce, R. Duncan and Raiffa, Howard. 1957. Games and Decisions. New York: Wiley.

Nash, John F. 1950a. The Bargaining Problem. Econometrica 18: 155 - 62 (April).

_____. 1950b. Equilibrium Points in N-Person Games. Proc. Nat. Acad. Sci. U.S.A. 36: 48 - 9.

_____. 1953. Two-Person Cooperative Games. Econometrica 21: 128 - 40(Jan).

Radner, Roy. 1980. Collusive Behavior in Noncooperative Epsilon-Equilibria of Oligopolies with Long but Finite Lives. Journ Econ. Theory 22: 22:136 - 54 (Apr).

Rosen, J. B. 1965. Existence and Uniqueness of Equilibrium Points for Concave N-Person Games. Econometrica 33: 520 - 34 (July).

Smale, Steve. 1980. The Prisoner's Dilemma and Systems Associated to Non-Cooperative Games. Econometrica 48: 1617 - 34 (Nov).

ARCC - Western Libraries - University of Western Ontario
D M SIDE SECT SHELF TRAY ITEM C
8 1 07 06 10 09 002 4

Telser, Lester G. 1972. Competition, Collusion and Game Theory. Chicago: Aldine.

_____ 1978. Economic Theory and the Core. Chicago: Univ of Chicago Press.

_____ 1980. A Theory of Self-Enforcing Agreements. Journ. of Bus. 53: 27 - 44 (Jan).