JOURNAL OF ECONOMIC THEORY 42, 13–29 (1987)

Coalition-Proof Nash Equilibria
II. Applications*

B. Douglas Bernheim

Stanford University, National Bureau of Economic Research,
Stanford, California 94305

AND

Michael D. Whinston

Harvard University, Cambridge, Massachusetts 02138

Received May 11, 1984; revised April 29, 1986

In Bernheim, Peleg, and Whinston ("Coalition-Proof Equilibria. I. Concepts," *J. Econ. Theory* **42** (1987), 1–12), we proposed the notion of Coalition-Proof Nash equilibrium and Perfectly Coalition-Proof Nash equilibrium as solution concepts for strategic environments in which players can freely discuss their strategies, but cannot make binding commitments. This paper undertakes applications to several economic problems, including the behavior of Cournot oligopolists, oligopolistic entry deterrence, cooperation in finite horizon games, and social choice rule implementation. *Journal of Economic Literature* Classification Numbers: 022, 025. © 1987 Academic Press, Inc.

In Bernheim, Peleg, and Whinston [5], we proposed the notion of Coalition-Proof Nash equilibrium as a solution concept for strategic environments in which players can freely discuss their strategies, but cannot make binding commitments. In addition, we introduced the idea of Perfectly Coalition-Proof Nash equilibrium, which also requires dynamic consistency of agreements reached by players. While we were unable to establish general conditions for the existence of such equilibria, we asserted that these concepts would prove useful in a variety of applied contexts.

In previous work (Bernheim and Whinston [6]), we developed one such application in detail. Specifically, we considered a rich class of auctions, in which bidders name "menus" of offers for various possible actions (allocations) available to the auctioneer. Our analysis was limited to situations in which bidders are symetrically informed. For such auctions,

* We would like to thank R. Aumann, F. Fisher, A. Mas-Colell, E. Maskin, and R. Myerson for helpful comments. The second author gratefully acknowledges support from the Sloan Foundation, through its grant to the MIT Department of Economics.

0022-0531/87 $3.00

Nash equilibria need not yield efficient allocations. However, we demonstrated that coalition-proof equilibria necessarily exist, are efficient, and yield the same payoffs that would arise if, in equilibrium, each bidder selected a menu offer which "truthfully" reflected his relative willingness-to-pay for various alternatives.

This paper undertakes applications of the coalition-proof and perfectly coalition-proof concepts to several other important economic problems. In Section 1, we consider the standard Cournot oligopoly model, and demonstrate that the unique Nash equilibrium is coalition-proof. In Section 2, we discuss the problem of oligopolistic entry deterrence, and argue that coalition-proofness helps to capture the intuitively appealing notion that incumbents may "gang up" on an entrant. In Section 3, we consider finite repetitions of a simple two player game, and demonstrate that all perfectly coalition-proof equilibria may involve cyclical cooperation, where the frequency of the cycle increases as the players' discount factors approch unity. Finally, in Section 4, we present some preliminary results concerning social choice rule implementation in coalition-proof equilibria.

1. A Static Cournot Oligopoly Model

Consider the standard n firm symmetric Cournot oligopoly problem. Each firm i produces some level of output, q_i, at cost $c(q_i)$. We assume that $c(\cdot)$ is convex. Market price is determined as a function of total output: $P(\sum_{i=1}^{n} q_i)$. We assume that $P(\cdot)$ is decreasing and concave.

Given the (possibly random) strategies of its competitors, firm i seeks to maximize expected profits (firms are risk neutral). Thus in equilibrium, any output level that i is willing to produce with positive probability must satisfy

$$\max_{q_i} E_i P\left(q_i + \sum_{j \neq i} q_j\right) q_i - c(q_i),$$

where E_{-i} is an operator which takes expectations over $(q_j)_{j \neq i}$. Clearly, i's objective function is strictly concave, so the solution to this problem is unique. As a result, we may, without loss of generality, confine attention to pure strategy equilibria.

It is well known that the game described above has a unique pure strategy Nash equilibrium (cf. Burger [7]). Furthermore, if we fix actions for any subset of players, the game induced upon its complement possesses the same essential features as the original symmetric Cournot oligopoly problem, and therefore also has a unique Nash equilibrium. Thus, there is no room for credible cooperation by any coalition of players. These arguments establish that the unique Nash equilibrium is coalition-proof.

It is interesting to note that this Coalition-Proof Nash equilibrium is *not* strong; indeed, no strong equilibrium exists for this model, specifically because this concept permits coalitions to deviate in ways which are *not* credible.

2. Oligopolistic Entry Deterrence

Students of industrial organization have long recognized that an incumbent firm may act to prevent the entry of new rivals into its market. In the model of Spence [18], for example, the incumbent undertakes preemptive investment that makes entry by new firms unprofitable.

In markets with several incumbents, however, deterrence expenditures may be worthwhile as a joint enterprise but not as an individual undertaking. In such cases, Perfect Nash equilibria both with and without entry can arise. Yet, if opportunities for communication exist and deterrence is mutually desirable, then the "entry" equilibrium requires the incumbents to ignore a mutually beneficial and self-enforcing opportunity to deter entry, which seems implausible. In such circumstances, only the "no entry" equilibrium is perfectly coalition-proof. In order to illustrate these and related points, we now consider a simple model of oligopolistic entry deterrence.

Consider a homogeneous product market with two incumbent firms, A and B, and a single potential entrant, E. The game proceeds as follows. First, each incumbent decides whether or not to undertake a cost reducing investment (we can also think of this investment as the construction of excess capacity). The investment costs γ and reduces the (constant) marginal costs of the firm from an initial level of c to zero. After observing the investment decisions of the incumbents, the potential entrant (E) decides whether or not to enter. E faces no variable production costs, but must incur a set-up cost of K upon entry. After the (publically observable) entry decision, operating firms simultaneously choose quantities (i.e., they play Cournot). Output price is given by $P(Q) = a - bQ$, where Q is total output. We assume no discounting of future profits.

This model gives rise to perfect equilibria in which firm E enters, as well as perfect equilibria in which the incumbents jointly deter entry. Our refinement has its greatest potential value when both types of equilibria coexist; we focus our discussion on these cases. The following four conditions are sufficient (and "almost" necessary) to guarantee the simultaneous existence of "entry" and "joint deterrence" equilibria:

$$\frac{a^2}{16b} < K, \tag{A.1}$$

$$\frac{(a+c)^2}{16b} > K, \tag{A.2}$$

$$\left(\frac{a-2c}{16b}\right)^2 \geqq \frac{(a+c)^2}{16b} - \gamma, \tag{A.3}$$

$$\frac{a^2}{9b} - \gamma \geqq \frac{(a-3c)^2}{16b}, \tag{A.4}$$

Condition (A.1) implies that investment by both incumbents succesfully deters entry. Written as a weak inequality (hence "almost"), (A.1) is plainly necessary for the existence of a perfect equilibrium without entry. Condition (A.2) implies that investment by one incumbent does not deter entry. It is not difficult to verify that (A.2), written as a weak inequality, is necessary for the coexistence of both types of equilibria.[1] Condition (A.3) indicates that neither incumbent will find it worthwhile to invest if the other does not invest. Given (A.1) and (A.2), a perfect Nash equilibrium with entry exists if and only if (A.3) holds. Condition (A.4), on the other hand, implies that each incumbent prefers to invest if he expects the other incumbent to do so as well. Given (A.1) and (A.2), joint deterrence arises in a Perfect Nash equilibrium if and only if (A.4) holds. It is possible to select K and γ to satisfy all four inequalities as long as $a > c(\frac{54}{7})^{1/2}$.[2]

When there exists Perfect Nash equilibria both with entry and with joint deterrence, incumbents have an incentive to coordinate actions. In particular, both A and B prefer the equilibrium with deterrence if and only if[3]

$$\frac{a^2}{9b} - \gamma > \frac{(a-2c)^2}{16b} \tag{A.5}$$

Since post-investment play is uniquely determined, we may replace subgames with the associated continuation payoffs, producing a static "investment game" between the incumbents. The perfectly coalition-proof concept then selects the Pareto efficient investment equilibrium. Thus, when both types of perfect equilibria coexist, only the one preferred by the incumbents is perfectly coalition-proof. This pattern accords with our

[1] Suppose that $(a+c)^2/16b < K$. Entry occurs in perfect Nash equilibrium only if (1) neither firm invests, and (2) neither firms wishes to invest unilaterally. The second condition holds if $(a-2c)^2/16b \geqq (a+c)^2/9b - \gamma$. For investment by both firms to be an equilibrium, we must have $a^2/9b - \gamma \geqq (a-2c)^2/9b$. Clearly, these two conditions cannot both hold.

[2] It is always possible to choose K to satisfy (A.1) and (A.2). On the other hand, we can choose γ to satisfy (A.3) and (A.4) only if $a \geqq c(\frac{54}{7})^{1/2}$.

[3] Note that there exist parameter values satisfying (A.1)–(A.4) under which incumbents prefer the equilibrium with deterrence (condition (A.5)), as well as parameter values satisfying (A.1)–(A.4) under which incumbents prefer the equilibrium with entry (not (A.5)).

intuition: the incumbents should not accept an inferior outcome which they could avoid through self-enforcing coordination in the first stage.

Finally, note that alternative refinements fail to isolate the appealing equilibrium. First, the Pareto criterion allows us to choose between the two perfect equilibria described above *only* when (A.5) fails to hold. It does *not* predict that the incumbents will necessarily coordinate their actions when it is in their mutual advantage to jointly deter entry (E strictly prefers the equilibrium with entry). Second, since equilibria are always Pareto inefficient (due to the lack of cooperation in output choices), neither Strong Nash nor Strong Perfect Nash equilibria exist. Finally, one could not obtain the desired outcome by simply applying the Coalition-Proof Nash concept to the normal form of this game, since this would permit equilibria which are sustained by dynamically inconsistent threats (e.g., E could threaten to enter and drive prices to zero if either incumbent invests).

3. Cyclical Cooperation in a Finitely Repeated Game

Several recent papers have addressed the problem of sustaining cooperation in finitely repeated games of complete information (cf., Radner [15], Benoit and Krishna [4], and Friedman [9]). As is well known, every game possessing a unique Nash equilibrium gives rise to a unique subgame perfect equilibrium in finite repetitions, consisting of the repeated static solution—attempts to establish collusion "unravel" from the terminal period. However, games possessing a multiplicity of Nash equilibria give rise to a host of other outcomes; indeed, Benoit and Krishna [4] establish under relatively weak conditions that, as in the infinite horizon case, every individually rational outcome can be obtained in some subgame perfect equilibrium. Of course, certain outcomes are often supported by punishment phases which, while subgame perfect, penalize all participants. When players have opportunities to communicate throughout the game, these punishments require the group to behave in a dynamically inconsistent fashion. Thus, the notion of perfectly coalition-proof equilibria may allow us to significantly refine the set of subgame perfect equilibria.

In Bernheim *et al.* [5], we presented a simple two period repeated game for which the unique perfectly coalition-proof equilibrium entailed repetitions of a single static solution, despite the existence of a Perfect Nash equilibrium that involves cooperation in the first period. This does not, however, reflect a general principle. Our current objective is to analyze the structure of perfectly coalition-proof equilibria in a finitely repeated game where participants can sustain cooperation.

We consider a simple two player game, which one can think of as a modification of the traditional prisoners' dilemma. Each player has four strategies: refuse to confess (N), or confess to one of three versions of the crime (A, B, C). Version A (C) implicates prisoner 1 (2) as the primary culprit, while version B implicates both equally. If both prisoners remain silent, the penalty is relatively mild. If only one confesses, the other receives a lengthy sentence. If both confess to the same version, the court imposes moderate punishments. If the prisoners give conflicting confessions, both are penalized heavily. Specific payoffs are given in Table I.

Now suppose that we repeat this game a finite number (K) of times. We suppose, in addition, that participants employ a common discount factor, ρ. Throughout, we confine attention to pure strategies in order to avoid the conceptual issue of whether, in each repetition, players observe each other's selection criteria, or simply the outcomes of randomization.

Is it possible to sustain collusion in subgame perfect equilibria? The answer to this question depends upon ρ. To enforce (N, N), one prescribes equilibrium actions of (B, B) for some specified number of terminal periods, and punishes deviations from collusion in previous periods by shifting to either (A, A) or (C, C). If $\rho \leqq \frac{1}{2}$, this punishment is insufficient to deter deviations from (N, N) in any period. However, if $\rho > \frac{1}{2}$, it will be possible to sustain cooperation for all but a fixed number (depending on ρ) of terminal periods.

In general, for any particular value of ρ, the set of subgame perfect equilibrium outcomes is very large. Some equilibria entail cooperation in almost every period, while some yield inefficient outcomes for every repetition. Trivially, one can construct cyclical equilibria, where play alternates between (N, N), (A, A), (B, B), and (C, C).

Only those outcomes which are efficient in the class of equilibria are coalition-proof (there are two players). However, in this context, it is more appropriate to refine the set of solutions by restricting attention to *perfectly* coalition-proof equilibria. We will now show that, for this example, perfectly coalition-proof equilibria have a peculiar and striking property: for identifiable interest rates, play *necessarily* cycles between cooperation and a

TABLE I

A Modified Prisoners' Dilemma

		Player 2			
		Not Confess	Confess A	Confess B	Confess C
Player 1	Not Confess	−1, −1	−11, 0	−10, 0	−9, 0
	Confess A	0, −9	−6, −4	−10, −10	−10, −10
	Confess B	0, −10	−10, −10	−5, −5	−10, −10
	Confess C	0, −11	−10, −10	−10, −10	−4, −6

static noncooperative outcome. In fact, one may induce cyclical cooperation of *any periodicity* as the unique consequence of perfectly coalition-proof behavior by selecting the discount factor appropiately. Further, the periodicity increases as the discount factor rises; when $\rho = 1$, players alternate between periods of cooperation and non-cooperation.

Our result employs the following notation. Let Z_+ be the set of non-negative integers. Let

$$[K/T] = \max\{J \in Z_+ \mid T \cdot J \leqq K\}$$

and

$$R(K, T) = K - T[K/T].$$

In addition, let ρ_τ be the value of ρ that solves

$$\frac{\rho}{1-\rho}(1 - \rho^{\tau-1}) = 1$$

for each $\tau \geqq 2$. Note that $\langle \rho_\tau \rangle_{\tau=2}^{\infty}$ is a decreasing sequence, with $\rho_2 = 1$ and $\lim_{\tau \to \infty} \rho_\tau = \frac{1}{2}$. We now state the result.

PROPOSITION 1. *Consider the K-repeated "modified prisoners' dilemma" game of Table* 1. *For all* $T \geqq 2$, *if the actual discount factor is* ρ_T, *then every perfectly coalition-proof equilibrium has the following features*[4]:

(i) *In the first* $R(K, T)$ *periods, players choose non-cooperative strategies* ((A, A), (B, B), *or* (C, C)) *in equilibrium*;

(ii) *Thereafter, equilibrium play cycles*; *each cycle consists of one period in which players cooperate* (*i.e., play* (N, N)) *followed by* $T-1$ *periods in which they fail to cooperate* (*i.e., play* (B, B)).

Before proving this result, we make two remarks. First, the equilibria described in Proposition 1 are plainly cyclical. The first $R(K, T)$ periods are too short for a cycle, and any static solution may prevail during this interval. In the following period, the players cooperate. Collusion breaks down for the next $T-1$ periods (players choose (B, B)); in the Tth subsequent period, players cooperate once again. Thereafter, the cycle repeats. The intuition for this phenomenon is actually quite simple. Suppose we solve for equilibria recursively from the terminal period. Collusion is clearly impossible in period K. However, it may be sustainable in some prior

[4] We adopt the convention that temporally earlier periods have lower numbers; the final period is period K.

period. Suppose it is sustainable for the first time in period $K-(T-1)$. Then, depending on payoffs, the collusive equilibrium for this subgame may Pareto dominate all other equilibria. If so, there is no room to punish defections in period $K-T$—each player knows that they would jointly defect to the mutually advantageous outcome in $K-(T-1)$. Thus, period $K-T$ effectively becomes a "terminal" period, and the process repeats itself.

Second, one naturally wonders about the nature of behaviour when the discount factor does not equal one of the critical levels described in Proposition 3. In particular, cyclical cooperation is *not* a "knife edge" phenomenon—one can show that perfectly coalition-proof equilibria necessarily involve cyclical cooperation with periodicity T for all interest rates in an interval $[\rho_T, \rho_T+\varepsilon_T]$, for some $\varepsilon_T > 0$. For $\rho \in (\rho_T+\varepsilon_T, \rho_{T-1})$, other possibilities may arise. Typically, these equilibria will also involve cycles, but the periodicity of the cycles may vary over time. It may, in fact, be possible to select discount rates which give rise to apparently chaotic cooperation (we have not yet found examples of this).

We now prove the result.

Proof of Proposition 1. We proceed by backwards induction. In period K, the only possible equilibrium outcomes are (A, A), (B, B), and (C, C). Now suppose that these are the only possible outcomes in periods $K+1-t$ through K. We claim that if $t<T-1$, then these are the only possible outcomes in period $K-t$.

Let S^τ be the set of sequences of strategies $s^\tau = \langle s_k \rangle_{k=1}^{\tau}$ such that $s_k \in \{A, B, C\}$ for each k. Let $\Pi_i(s_k)$ be the payoffs associated with (s_k, s_k) for player i (i.e., if $s_k = A$, $\Pi_1(s_k) = -6$). Let

$$\Pi_i^\tau(s^\tau) = \sum_{k=0}^{\tau-1} \rho_T^k \Pi_i(s_k).$$

Finally, let $a^\tau = (A, A,..., A)$, $b^\tau = (B, B,..., B)$, $c^\tau = (C, C,..., C)$.

Now we know that in periods $K+1-t$ through K, equilibrium play can be represented by some sequence $s^t \varepsilon S^t$. The worste punishments which we can inflict on these players in periods $K+1-t$ through K yield losses (discounted to period $K-t$) of

$$p_1^t(s^t) = \rho_T(\Pi_1^t(s^t) - \Pi_1^t(a^t)),$$
$$p_2^t(s^t) = \rho_T(\Pi_2^t(s^t) - \Pi_2^t(c^t)),$$

to players 1 and 2, respectively. But note that

$$p_1^t(s^t) + p_2^t(s^t) = 2 \sum_{k=1}^{t} \rho_T^t$$

$$= 2 \frac{\rho_T}{1 - \rho_T} (1 - \rho_T^t)$$

$$< 2$$

since $t < T - 1$. But then $p_i^t(s^t) < 1$ for some i. For each i, the current gain from deviating is 1 or greater for every outcome other than (A, A), (B, B), or (C, C). Thus, in period $K - t$, we must have one of these outcomes. Further, since total payoffs are the same in all three cases, we cannot rule out any of these outcomes through Pareto dominance.

Now consider $t = T - 1$. Clearly, for all $s^{T-1} \in S^{T-1}$, and $i = 1, 2$,

$$p_i^{T-1}(s^{T-1}) \leqq 2$$

For each i, the current gain from deviating is 4 or greater for every outcome other than (A, A), (B, B), (C, C), or (N, N). Thus, in period $K + 1 - T$, we must have one of these outcomes.

In particular, we may have (N, N).

$$p_i^{T-1}(b^{T-1}) = 1$$

for $i = 1, 2$, and the current gain from deviating is 1 for $i = 1, 2$ when the outcome is (N, N). Further, it is easy to check that in period $K + T - 1$, (N, N) can be enforced only if it is followed by b^{T-1} on the equilibrium path.

Thus, from period $K + 1 - T$ onwards, we have several possible equilibrium outcomes: (N, N) followed by b^{T-1}, or an element of S^T. We now argue that the first of these dominates all of the others:

$$\Pi_i(N) + \rho_T \Pi_i^{T-1}(b^{T-1}) = -6,$$

while, for all $s^T \in S^T$,

$$\Pi_i^T(s^T) \leqq -4 - 4\rho_T \sum_{k=0}^{T-2} \rho_T^k = -8.$$

Thus, from period $K + 1 - T$ onwards, the only possible perfectly coalition-proof equilibrium outcome is (N, N) followed by b^{T-1}.

Now consider period $K - T$. Since the outcome in subsequent periods is uniquely determined, there is no ability to punish current deviations. This period is, effectively, equivalent to the terminal period K. Thus, we repeat the above analysis. The proposition is established by induction on the number of cycles. Q.E.D.

Given the highly specific nature of our example, one naturally wonders about the generality of the properties described in Proposition 1. Indeed, it is a simple matter to design finitely repeated games in which cyclical behavior does not necessarily arise in perfectly coalition-proof equilibria for any discount factor. We have not yet obtained an interesting set of conditions characterizing cyclicity; we leave this task for future work.

4. Social Choice Rule Implementation

A social choice rule is a correspondence which selects a set of optimal social states for each possible configuration of individual preferences. The difficulties involved in constructing a social choice rule which satisfies various potentially desirable properties have been well known since Arrow's pathbreaking theorem [1].[5] Over the past fifteen years a large literature has emerged which deals with a related problem—that of "implementing" any given social choice rule.[6] This problem arises whenever the social planner lacks information concerning individuals' preference relations. Rather than dictate a social optimum, the planner must instead design a mechanism for social decision making (a "game form") which provides participants with appropriate incentives, thereby generating optimal outcomes. This literature has sought necessary and sufficient conditions under which any particular social choice rule is implemented by *some* game form.

To be more specific, let A be the set of feasible social states, and let $\mathbf{R}_i$ denote individual i's set of possible weak preference relations over A. We use R_i to denote an element $\mathbf{R}_i$ (P_i indicates the strong preference relation associated with R_i). Let $\mathbf{R}=\Pi_{i=1}^{n}\mathbf{R}_i$. A social choice rule is a correspondence $f\colon \mathbf{R}\rightrightarrows A$, which indicates optimal social states for each profile of preferences. A game form for players $\{1,..., n\}\equiv N$ is a collection of strategy sets $\{S^i\}_{i=1}^{n}$, and an outcome function, $g\colon S\rightarrow A$ (where $S=\prod_{i=1}^{n} S_i$). Let Γ denote the set of possible game forms. An equilibrium concept is a correspondence $E\colon \Gamma\times\mathbf{R}\rightrightarrows A$ which indicates equilibrium solutions for each game form and preference profile. We now formally define the notion of "implementation":

Definition. A social choice rule is said to be *fully implemented* under equilibrium concept E by a particular game form g if and only if for all $R\in\mathbf{R}$, $f(R)=E(g, R)$. A social choice rule is said to be *implementable*

[5] Arrow's theorem concerned social preference orderings, which provide more information than social choice rules.

[6] See, for example, the discussion and references in Dasgupta, Hammond, and Maskin [8].

under equilibrium concept E if and only if it is fully implemented by *some* $g \varepsilon \Gamma$.

Maskin has derived necessary and sufficient conditions for implementability in both Nash [10] and Strong Nash [11] equilibria (see also his survey article [12]). For both concepts, he shows that implementable social choice rules are necessarily "monotonic" (following Muller and Satterthwaite [13], we will refer to this property as "strong positive association"—see below). In addition, he proves [11, 12] that when $|A| \geqq n \geqq 3$, no social choice rule (defined on unrestricted preference domains) satisfying "no veto power" is implementable in Strong Nash equilibria. Finally, he demonstrates [10, 12] that any social choice rule satisfying both monotonicity and no veto power is implementable in Nash equilibria.

Here we examine the problem of social choice rule implementation in Coalition-Proof Nash equilibria.[7, 8] We begin our analysis by showing that for this solution concept, implementable social choice rules necessarily satisfy "lower strong positive association" (see Barbera and Dutta [3]). This condition is *weaker* than strong positive association. Further, we demonstrate, by way of example, that strong positive association is *not* a necessary condition for implementability under the coalition-proof concept. This result is somewhat surprising given Maskin's results and the relationships between the Nash, Strong Nash, and Coalition-Proof Nash equilibrium sets of a game. One might, for example, think that implementability in Coalition-Proof Nash equilibria implies implementability in Nash equilibria, since all Coalition-Proof Nash equilibria are Nash equilibria. However, this is not the case. Recall that the notion of implementation requires the set of equilibrium states and socially preferred states to be *equivalent*. When a social choice role is implementable in Coalition-Proof Nash equilibria, it may be that the Nash equilibrium set is too large. In addition, our example also establishes that $|A| \geqq n \geqq 3$ and the no veto power property do *not* preclude a choice rule from being implementable in coalition-proof equilibria. Finally, we provide a set of sufficient conditions for implementability. Specifically, under a certain natural restriction of the preference domain, any Pareto optimal social choice rule that is implementable in Nash equilibria is also implementable in Coalition-Proof Nash equilibria.

[7] Note that we are *not* using the Perfectly Coalition-Proof Nash equilibrium concept here. It is easy to see that any social choice rule that is implementable under the coalition-proof concept is also implementable under the perfectly coalition-proof concept. The converse, however, is not true.

[8] Peleg [14] also investigates social choice rule implementation in Coalition-Proof Nash equilibrium although he presses in a somewhat different direction than that persued here.

We begin by introducing the following definitions, which we interpret below.

DEFINITION. $f\colon \mathbf{R} \rightrightarrows A$ satisfies *strong positive association* (SPA) iff for all R, $R' \in \mathbf{R}$, $a \in f(R)$ and [for all $i \in N$ and for all $b \in A$, $aR_ib \Rightarrow aR'_ib$] imply $a \in f(R')$.

DEFINITION. $f\colon \mathbf{R} \rightrightarrows A$ satisfies *lower strong positive association* (LSPA) iff for all R, $R' \in \mathbf{R}$, $a \in f(R)$ and [for all $i \in N$ and for all $b, c, d \in A - \{a\}$ where cR_ia, $cR_id \Leftrightarrow cR'_id$, $cP_id \Leftrightarrow cP'_id$, $aR_ib \Rightarrow aR'_ib$, and $aP_ib \Rightarrow aP'_ib$] imply $a \in f(R')$.

DEFINITION. $f\colon \mathbf{R} \rightrightarrows A$ satisfies *no veto power* (NVP) iff, for all $R \in \mathbf{R}$ and for all $a \in A$, if there exists $j \in N$ such that for all $i \neq j$, aR_ib for all b, then $a \in f(R)$.

DEFINITION. $f\colon \mathbf{R} \rightrightarrows A$ satisfies *Pareto optimality* (PO) iff, for all $R \in \mathbf{R}$ and for all $b \in A$, if there exists $a \in A$ such that aP_ib for all i, then $b \notin f(R)$.

Strong positive association requires the following. Take a profile of preferences (R) under which a is chosen. Reorder the alternatives within the sets $\{b \mid bP_ia\}$ and $\{c \mid aR_ic, c \neq a\}$, and then either move alternative a up in i's ordering, or leave its position unchanged. Alternative a must still be chosen. SPA is clearly related to Arrow's "independence of irrelevant alternatives," and therefore rules out comparisons of preference intensity (in particular, if we *just* rearrange alternatives within the sets indicated above without moving alternative a at all, then a is still chosen). For example, the Borda rule violates SPA. Lower strong positive association is weaker than SPA. LSPA requires that alternative a must remain optimal when we change each i's preferences by reordering within the set $\{b \mid aP_ib\}$, and then either move a up, or leave its position unchanged in i's ordering—we do not alter an individual's ordering within the set $\{c \mid cR_ia, c \neq a\}$. Plurality rule (a special case of the Borda rule, in which each individual's highest ranked alternative receives one "point") satisfies LSPA, but not SPA. The definitions of "no veto power" and "Pareto optimality" are standard, and more straightforward.

We begin our analusis of necessary conditions for implementability with

LEMMA 1. *Consider any game form,* $(\{S_i\}_{i=1}^n, g)$. *Let* $s^* \in S$ *with* $g(s^*) = a$ *be a Coalition-Proof Nash equilibrium when preferences are* R. *Let* R' *be such that* [*for all* i, *and for all* $b, c, d \in A - \{a\}$, *where* cR_ia, $cR_id \Leftrightarrow cR'_id$, $cP_id \Leftrightarrow cP'_id$, $aR_ib \Rightarrow aR'_ib$, *and* $aP_ib \Rightarrow aP'_ib$]. *Then* s^* *is a Coalition-Proof Nash equilibrium when preferences are* R'.

Proof. First, the lemma is obviously true for $n=1$. Now assume the lemma is true for all $n=1,\dots,m-1$. We show that the lemma is also true for $n=m$. Suppose not, i.e., that s^* is not a Coalition-Proof Nash equilibrium under preferences R'. Then, either s^* is not self-enforcing under R' or, under R', there exists another self-enforcing Nash equilibrium, $\bar{s}$, which Pareto dominates s^*. Suppose the former is true. Then there exists a $J \subset \{1,\dots,m\}$ and an s_J which is a Coalition-Proof Nash equilibrium in the game induced on J by s^*_{-J} and is such that $g(s_J, s^*_{-J})\, P'_i\, g(s^*)$ for all $i \in J$.[9] But, note that relative to $g(s_J, s^*_{-J})$ the movement from R'_J to R_J (the restrictions of R' and R to $i \in J$) satisfies the hypothesis of the lemma. Thus, since we assume that the lemma is true for $n<m$, s_J is a Coalition-Proof Nash equilibrium in the game induced on J by s^*_{-J} under preferences R_J also. But, since $g(s_J, s^*_{-J})\, P'_i\, g(s^*)$ implies $g(s_J, s^*_{-J})\, P_i\, g(s^*)$, we have a contradiction to the hypothesis that s^* is a Coalition-Proof Nash equilibrium under preferences R.

Now suppose that under R', there exists another self-enforcing equilibrium, $\bar{s}$, which Pareto dominates s^*. Then $\bar{s}$ is a self-enforcing equilibrium such that $g(\bar{s})\, P'_i\, g(s^*)$ for all i. Clearly, this implies $g(\bar{s})\, P_i\, g(s^*)$ for all i, so that if $\bar{s}$ is also self-enforcing under R we have a contradiction to the hypothesis that s^* is coalition-proof under R. Suppose, instead, that $\bar{s}$ is not self-enforcing under R. Then there exists a $K \subset \{1,\dots,m\}$ and an s_K such that s_K is a Coalition-Proof Nash equilibrium in the game induced on K by $\bar{s}_{-K}$ (when preferences are R) and $g(s_K, \bar{s}_{-K})\, P_i\, g(\bar{s})$ for all $i \in K$. Then, since $g(s_K, \bar{s}_{-K})\, P_i\, g(\bar{s})\, P_i\, g(s^*)$ and $g(\bar{s})\, P'_i\, g(s^*)$ for all $i \in K$, the movement from R_K to R'_K also satisfies the hypotheses of the lemma (taking $a = g(s_K, \bar{s}_{-K})$), so that s_K is also a Coalition-Proof Nash equilibrium in the game induced on K by $\bar{s}_{-K}$ when preferences are R'_K. But this implies that $\bar{s}$ is not self-enforcing under R'—a contradiction. Thus, the lemma is indeed true for $n=m$. Apply induction.

Q.E.D.

We now state and prove our fundamental result, which establishes a necessary condition for implementability under the Coalition-Proof Nash equilibrium concept.[10]

PROPOSITION 2. *If $f\colon \mathbf{R} \to A$ is implementable in Coalition-Proof Nash equilibria, then f satisfies lower strong positive association* (LSPA).

Proof. Let the game form $[g, \{S^i\}_{i=1}^n]$ fully implement f. Suppose that f does not satisfy (LSPA). Then there exist R and R' and an $a \in f(R)$ such

[9] Actually, there must exist a self-enforcing equilibrium for the players in J which Pareto dominates s^*—but this implies that there exists a coalition-proof equilibrium that does so as well.

[10] Peleg [14] also proves a slightly weaker version of this result.

that [for all i, and for all $b, c, d \in A - \{a\}$ where $cR_i a$, $cR_i d \Leftrightarrow cR_i' d$, $cP_i d \Leftrightarrow cP_i' d$, $aR_i b \Rightarrow aR_i' b$, $aP_i b \Rightarrow aP_i' b$], but $a \notin f(R')$. Since f is fully implemented by this game form, there exists an $s^* \in S$ such that $g(s^*) = a$ and s^* is a Coalition-Proof Nash equilibrium when preferences are R. By Lemma 1, s^* is also a Coalition-Proof Nash equilibrium when preferences are R'. Since $a \notin f(R')$, g cannot fully implement f—a contradiction. Q.E.D.

We now demonstrate by way of example that SPA is not a necessary condition for implementability in coalition-proof equilibria (this contrasts with Maskin's results for the Nash and Strong Nash concepts). In addition, this example illustrates that NVP (and $|A| \geqq n \geqq 3$) does not preclude implementability in coalition-proof equilibria (again, this contrasts with Maskin's result for the Strong Nash concept).

Consider a setting in which there are three individuals ($n = 3$). For all i, let $\mathbf{R}_i$ consist of all possible strong orderings over A and, for any set of preference orderings R, let $z_R(i)$ be individual i's most preferred alternative. Define the plurality choice rule, f^p, by

$$f^p(R) = \begin{cases} \{x\} & \text{if } x = z_R(i) = z_R(j) \text{ for some } i \neq j \\ A & \text{if } z_R(i) \neq z_R(j) \text{ for all } i \neq j. \end{cases}$$

It is easy to verify that f^p satisfies NVP. To see that f^p violates SPA, consider the following two sets of preference profiles (where $A = \{a, b, c\}$):

R_1	R_2	R_3	R_1'	R_2'	R_3'
a	b	c	a	c	c
b	c	a	b	a	a
c	a	b	c	b	b

Note that $f^p(R) = \{a, b, c\}$ and $f^p(R') = c$. Clearly $aR_i x \rightarrow aR_i' x$ for all i and for all $x \in A$, so that f^p does not satisfy SPA. We now show that f^p is, nevertheless, implementable in coalition-proof equilibria.

Proposition 3. *f^p is implementable in Coalition-Proof Nash equilibria.*

Proof. Define $L(a, R_j) \equiv \{b \in A \mid aR_j b\}$. Maskin [10] constructs a game form, G_n^*, with the following four properties (we describe this game form in the Appendix):

(1) $S^i = \{(R, a) \mid R \in \mathbf{R}, a \in A\}$ for all $i \in N$.

(2) If $s_i = (R, a)$ for all $i \in N$ and $a \in f(R)$, then $g(s) = a$.

(3) If $s_i = (R, a)$ for all $i \neq j$ and $a \in f(R)$, then $\{b \in A \mid b = g(s_j, s_{-j}), s_j \in S^j\} = L(a, R_j)$.

(4) If either: (i) $s_i \neq s_k$ for some $i, k \neq j$ or (ii) $s_i = (R, a)$ for all $i \neq j$ but $a \notin f(R)$, then $\{b \mid b = g(s_j, s_{-j}), s_j \in S_j\} = A$.

It is now fairly straightforward to confirm that $G_3^\ddagger$ (the special case of this game in which $n = 3$) fully implements f^p in coalition-proof equilibria. First, we argue that if $a \in f^p(R)$, then the strategy configuration $s_i = (R, a)$ for all i is a Coalition-Proof Nash equilibrium. Since f^p satisfies Pareto optimality, we need only show that this strategy profile is self-enforcing. By property (3), no unilateral deviation is profitable. We now consider two-player joint deviations. To see that these cannot occur, note that such a deviation would yield a strategy profile for which property (4) would apply for both deviating players. Thus, unless the two players agree upon their favorite outcome, this deviation cannot be self-enforcing. But if they do agree on their favorite outcome then, by definition of f^p, this must be outcome a. Thus, no mutually beneficial and self-enforcing deviation exists.

We now show that when preferences are R, no non-f^p-optimal outcome can arise as a Coalition-Proof Nash equilibrium. If outcome b arises and $b \notin f^p(R)$, then there exist two players, i and j, such that $z_R(i) = z_R(j) \neq b$. By property (4), this can only be a Nash equilibrium if $s_1 = s_2 = s_3$. But, again by property (4), if $s_1 = s_2 = s_3$ there exists a self-enforcing, mutually beneficial opportunity to deviate for players i and j (property (4) implies that, starting at $s_1 = s_2 = s_3$, i and j can choose their strategies so as to attein any outcome in A, including $z_R(i) = z_R(j)$). Thus, no outcome $b \notin f^p(R)$ can arise as a Coalition-Proof Nash equilibrium when preferences are R. We conclude that $G_3^\ddagger$ fully implements f^p. Q.E.D.

Although plurality rule is implementable in coalition-proof equilibria when $n = 3$, this does not appear to be the case in larger populations. While interesting, this observation does not alter the central implications of Proposition 3, concerning the roles of SPA and NVP.

Our final result provides a set of conditions which guarantee that a social choice rule is implementable in coalition-proof equilibria. We show that if one restricts the preference domain to profiles for which no two individuals agree on the best outcome, any Pareto optimal social choice rule that is fully implementable in Nash equilibria is also fully implementable in Coalition-Proof Nash equilibria. While demanding, this domain restriction would be satisfied, for example, in any economic context for which the set A included the allocation of some private good.

The result requires some additional notation. Let

$$\mathbf{R}^* \equiv \{R \in \mathbf{R} \mid z_R(i) \neq z_R(j) \text{ for all } i \neq j\}$$

We establish:

PROPOSITION 4. *Suppose* $|A| \geqq n \geqq 3$. *Any social choice rule* $f: \mathbf{R}^* \to A$ *which satisfies* PO *and* SPA *is implementable in Coalition-Proof Nash equilibria.*

Proof. We shall show that the game form G_n^*, described in the preceding example, fully implements f if the hypotheses of the theorem hold.

We first argue that if $a \in f(R)$, then $s_i^* = (R, a)$ is a Coalition-Proof Nash equilibrium. First, note that if $(s_1^*, ..., s_n^*)$ is self-enforcing, then by PO it is coalition-proof. To see that it is self-enforcing, note first that, by property (2), it is immune to unilateral deviations. But, by property (4) and the fact that no two individuals agree on their favorite outcome, no coalitional deviation is self-enforcing. Thus, $(s_1^*, ..., s_n^*)$ is a Coalition-Proof Nash equilibrium.

Next, we show that no $b \notin f(R)$ can arise as a Coalition-Proof Nash equilibrium when preferences are R. First, since no two individuals agree on their favorite outcome, property (4) ensures that no strategy profile that does not have $s_i = (R', b)$ for all i and $b \in f(R')$ can be a Coalition-Proof Nash equilibrium (otherwise, profitable unilateral deviations would be present). Second, if $s_i = (R', b)$ with $b \in f(R')$ for all i does constitute a Coalition-Proof Nash equilibrium, then by property (3) it must be true that, for all $c \in A$, $bR_i'c \to bR_ic$ for all $i \in N$. But then, SPA implies that $b \in f(R)$. Thus, G_n^* does in fact fully implement f. Q.E.D.

APPENDIX

To define G_n^* we first write $A = \{a(0), ..., a(m-1)\}$. Next, define $m(a, R_j) = |L(a, R_j)|$ and write

$$L(a, R_j) = \{b(0; a, R_j), b(1; a, R_j), ..., b(m(a, R_j) - 1; a, R_j)\}.$$

Maskin [10] specifies G_n^* as follows:

(1) $S^i = \{(R, a) | R \in \mathbf{R}, a \in A\}$ for all $i = 1, ..., n$.

(2) $g(s) = a$ if $s_1 = \cdots = s_n = (R, a)$ and $a \in f(R)$.

(3) If $s_i = (R_1, ..., R_n, a(r))$ for all $i \neq j$ where $a(r) \in f(R)$, and $s_j = (R', a(t)) \neq s_i$, then $g(s) = b([r + t] | m(a(r), R_j); a, R_j)$ where $[x] | y \equiv \min\{k \geqq 0 | x - k$ is a multiple of $y\}$.

(4) If $s = [(R^1, a(t_1)), ..., (R^n, a(t_n))]$ is such that there exists i, j, k such that $s_i \neq s_j \neq s_k \neq s_i$, then $g(s) = a([\sum_{q=1}^n t_q] | m)$.

(5) If $s_i = (R, a(r))$ for all $i \neq j$ where $a(r) \notin f(R)$, and $s_j = (R', a(t)) \neq s_i$, then $g(s) = a([r + t] | m)$.

REFERENCES

1. K. J. ARROW, "Social Choice and Individual Values," Yale Univ. Press, New Haven, CT., 1951.
2. R. AUMANN, Acceptable points in general cooperative *n*-person games, *in* "Contributions to the Theory of Games IV," Princeton Univ. Press, Princeton, N. J. 1959.
3. S. BARBERA AND B. DUTTA, Implementability via protective equilibria, *J. Math. Econom.* **10** (1982), 49–65.
4. J.-P. BENOIT AND V. KRISHNA, Finitely repeated games, *Econometrica* **53** (1985), 905–922.
5. B. D. BERNHEIM, B. PELEG, AND M. D. WHINSTON, Coalition-Proof Equilibria. I. Concepts, *J. Econ. Theory* **42** (1987), 1–12.
6. B. D. BERNHEIM AND M. D. WHINSTON, Menu auctions, resource allocation, and economics influence, *Quart. J. Econom.* **101** (1986), 1–31.
7. E. BURGER, "Introduction to the Theory of Games," Prentice–Hall, Englewood Cliffs, N. J., 1963.
8. P. DASGUPTA, P. HAMMOND, AND E. MASKIN, The implementation of social choice rules: Some general results on incentive compatibility, *Rev. Econom. Stud.* **46** (1979), 185–216.
9. J. FRIEDMAN, Cooperative equilibria in finite horizon noncooperative supergames, *J. Econ. Theory* **35** (1985), 390–398.
10. E. MASKIN, Nash equilibrium and welfare optimality, *in* "Mathematics of Operations Research," MIT mimeo, 1977, in press.
11. E. MASKIN, Implementation and strong Nash equilibrium, *in* "Aggregation and the Revelation of Preferences" (J. J. Laffont, Ed.), North-Holland, New York, 1979.
12. E. MASKIN, "The Theory of Implementation in Nash Equilibrium: A Survey," MIT Working Paper, No. 333, Reading, Mass., October 1983.
13. E. MULLER AND M. SATTERTHWAITE, The equivalence of strong positive association and strategy-proofness, *J. Econ. Theory* **14** (1977), 412–418.
14. B. PELEG, "Quasi-Coalitional Equilibria, Part I. Definitions and Preliminary Results," Center for Research in Mathematical Economics and Game Theory, The Hebrew University, Jerusalem, Research memorandum No. 59, 1984.
15. R. RADNER, Collusive behavior in noncooperative epsilon-equilibria of oligopolies with long but finite lives, *J. Econ. Theory* **22** (1980), 136–154.
16. A. RUBINSTEIN, Strong perfect equilibrium in supergames, *Internat. J. Game Theory* **9** (1980), 1–12.
17. R. SELTEN, A reexamination of the perfectness concept for equilibrium points in extensive games, *Internat. J. Game Theory* **4** (1975), 25–55.
18. A. M. SPENCE, Entry, investment, and oligopolistic pricing, *Bell J. Econom.* **8** (1977), 534–44.