GAMES AND ECONOMIC BEHAVIOR 3, 278–294 (1991)

Extending Renegotiation-Proofness to Infinite Horizon Games*

GEIR B. ASHEIM

The Norwegian School of Economics and Business Administration, N-5035 Bergen-Sandviken, Norway

Received December 5, 1988

This paper extends the concept of a *Pareto-perfect* equilibrium to infinite horizon games, and offers a definition that covers both finitely and infinitely repeated games. An equilibrium is said to be Pareto-perfect if admitted by a *Pareto-perfect social norm.* A Pareto-perfect norm admits an equilibrium if and only if it cannot be profitably renegotiated in a norm-observing manner at a future contingency. The defined concept is compared to concepts that other writers have proposed. It departs from other concepts by insisting that nonviable equilibria be defeated through renegotiation by viable ones, and by *not* imposing a stationarity assumption. *Journal of Economic Literature* Classification Number: 026. © 1991 Academic Press, Inc.

1. INTRODUCTION

This paper extends the concept of a *Pareto-perfect* equilibrium (PPE) to infinite horizon games, and offers a definition that covers both finitely and infinitely repeated games. The present analysis builds on the important contributions by Bernheim and Ray (1989) (henceforth referred to as B&R) and Farrell and Maskin (1989) (F&M), but obtains a concept with different implications.

* This research was initiated during fruitful visits to the Universities of Bonn and Haifa, and grew out of discussions with Joseph Greenberg based on his *Theory of Social Situations.* I thank a referee and, especially, an associate editor for very constructive criticism. Thanks also to seminar participants at Bielefeld, Cambridge, LSE, Munich, Ohio State, Oslo, and Tel Aviv for helpful comments. Support by Stifterverband für die Deutsche Wissenschaft (Ruhrgas-Stipendium) and by Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 303) is gratefully acknowledged.

0899-8256/91 $3.00

The underlying motivation for renegotiation-proofness is as follows: In an infinitely repeated game, a Subgame-perfect equilibrium (SPE) can be supported by the threat that a deviation by one player will trigger a grim continuation equilibrium unattractive to all players. However, allowing for renegotiation, the players will not submit to this grim equilibrium since they ex post favor returning to the original one. But then the threat supporting the original equilibrium is no longer credible. Hence, the question is: *what SPEa are not subject to this criticism*?

In finitely repeated games, a natural procedure is to determine viable continuation equilibria in the last stage by the Pareto-dominance refinement on the set of Nash equilibria, and then recursively determine viable continuation equilibria in earlier stages. This yields Pareto-perfect[1] equilibria (see Definition 1′ of Section 5).[2]

In infinitely repeated games, such a recursion cannot be used. Instead F&M propose the concept of a *Weakly renegotiation-proof* (WRP) equilibrium (independently suggested by B&R and called *Internally consistent* by them). A WRP equilibrium σ is an SPE which does not have two continuation equilibria such that the players all prefer moving from the one to the other. This concept takes care of *internal stability:* for no history can the players profitably renegotiate to an SPE in the set of continuation equilibria of σ. However, it lacks *external stability:* it is left completely unexplained why the players do not renegotiate to an SPE outside the set of continuation equilibria of σ. In order to impose external stability, F&M suggest the alternative concept of a *Strongly renegotiation-proof* (SRP) equilibrium (being a WRP equilibrium which does not have a continuation equilibrium from which the players all prefer moving to some other WRP equilibrium), while B&R define the related concept of a *Consistent* equilibrium (yielding existence also when WRP equilibria upset each other cyclically). There are, however, grounds to question the definitions of SRP and Consistent equilibria (reproduced in Section 3) on various accounts. In particular,

(1) these concepts may eliminiate SPEa that are defeated through renegotiation only by equilibria that are themselves not viable;

(2) being based on the concept of a WRP equilibrium, they impose a stationarity assumption, which may not be justified.

[1] This terminology is taken from Bernheim and Ray (1985) where the concept was first suggested. However, Selten's (1973) concept of *Payoff-optimality* is similar in spirit.

[2] For games with more than two players, it may be reasonable to assume that coalitions short of the grand coalition also can agree on coalitional deviations in some subgame. By a similar recursive method (see Bernheim *et al.*, 1987), this yields the concept of a *Perfectly coalition-proof* equilibrium in finite horizon games.

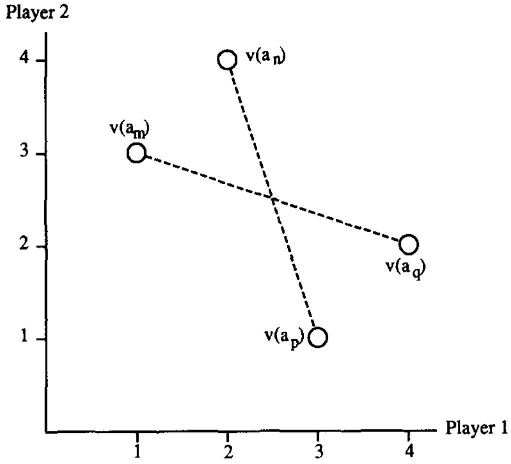

FIG. 1. Example 1.

The former point is illustrated by Example 1, shown in Fig. 1, and analyzed in Section 3 (where the payoff matrix is provided). In the figure, denote by $a_j, j = m, n, p, q$, action profiles and by v the payoff function of the stage game. Using Abreu's (1988) concept of a *simple strategy profile*,[3] the SPEa of interest are: $\sigma_n := \sigma(\pi_n, \pi_n, \pi_n)$, $\sigma_p := \sigma(\pi_p, \pi_n, \pi_n)$, and $\sigma_q := \sigma(\pi_q, \pi_m, \pi_m)$, where π_j for each j denotes the infinite repetition of a_j. Here, σ_n, σ_p, and σ_q are all WRP, while σ_n is SRP and Consistent since for no history is profitable renegotiation feasible. Furthermore, σ_q is not SRP nor Consistent since both players prefer σ_n to some continuation profile of σ_q. Finally, σ_p is not SRP since both players prefer σ_q to σ_p, while (informally) σ_p is not Consistent since σ_n indirectly upsets σ_p through σ_q.

Hence, according to the concepts of F&M and B&R, σ_n is renegotiation-proof, while σ_p is not. Note, however, that the SPE σ_p can be profitably renegotiated only to an SPE like σ_q which itself is nonviable when allowing for renegotiation (since after a deviation, the grand coalition will, rather than carrying out the punishment path π_m, renegotiate to σ_n). Hence, if one were to insist that a nonviable SPE be defeated through renegotiation in some subgame by a viable equilibrium, one would have to disagree with the conclusion that σ_p in nonviable. The concept of renegotiation-proofness proposed in this paper departs from F&M's SRP equi-

[3] A *simple strategy profile* $\sigma(\pi_0, \pi_1, \pi_2)$ is described by: (i) start with π_0, and (ii) if player $i \in \{1, 2\}$ is the unique deviatior from π_j (re)start π_i, otherwise continue with π_j; where π_j, $j \in 0, 1, 2$, are paths.

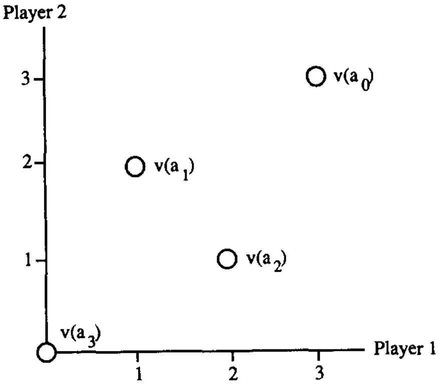

FIG. 2. Example 2.

librium and B&R's Consistent equilibrium especially through requiring that a nonviable SPE be defeated through renegotiation by a viable one. In particular, in Example 1 it leads to the result that σ_p is also a renegotiation-proof equilibrium (see Claim 1 of Section 3).

The fact that the stationarity assumption underlying the definitions of SRP/Consistent equilibria may not be warranted is illustrated by Example 2, shown in Fig. 2, and analyzed in Section 4 (where the payoff matrix is provided). The SPEa of interest are $\sigma_0 := \sigma(\pi_0, \pi_1, \pi_2)$, $\sigma_1 := \sigma(\pi_1, \pi_1, \pi_2)$, $\sigma_2 := \sigma(\pi_2, \pi_1, \pi_2)$, and $\sigma_3 := \sigma(\pi_3, \pi_3, \pi_3)$, where $\pi_j, j = 0, 3$, denote the infinite repetition of a_j, while $\pi_j, j = 1, 2$, denote the play of a_j once followed by π_0. Here all SPEa but the unanimously least attractive, viz. σ_3, have continuation equilibria which Pareto-dominate each other; i.e., they are not WRP and cannot be candidates for SRP/Consistent equilibria. Hence, these concepts uniquely determine as renegotiation-proof the SPE σ_3 which yields each player the lowest individually rational payoff, even though σ_0, σ_1, and σ_2 for every history Pareto-dominates σ_3. In other words, by ruling out the ability to inflict a mild punishment after a deviation (since it necessarily hurts all players and thereby leads to renegotiation), the concepts of F&M and B&R in this example predict that the grand coalition will in effect be forced to accept a *harsh* "punishment" at *every* contingency. This prediction does not seem too reasonable and is *not* shared by the concept of the present paper. In fact, Claim 2 of Section 4 establishes σ_0, σ_1, and σ_2 as renegotiation-proof, while σ_3 is not.

The concept of renegotiation-proofness presented in this paper can be described as follows: Consider a *social norm* that for any history admits some subset of the SPEa in the remaining subgame. Call the grand coalition *norm-observing* if for a given history it chooses an SPE admitted by the norm. Impose *internal stability* on such a norm by requiring that an SPE admitted by the norm cannot be profitably renegotiated by the grand coalition in a norm-observing manner at any future contingency. Impose *external stability* on such a norm by requiring that an SPE that violates the norm *can* be profitably renegotiated by the grand coalition in a norm-observing manner at some future contingency. A norm is called *Pareto-perfect* (PP) if it is both internally and externally stable. In games with a unique PP norm, a clear-cut definition of renegotiation-proofness is obtained; An SPE of a repeated game is *Pareto-perfect* if the unique PP norm admits this equilibrium given the initial history. In infinitely repeated games, however, multiple PP norms may occur. Then an SPE is said to be *Pareto-perfect* if at least one PP norm admits this equilibrium given the initial history.

In Example 1, there is a unique PP norm. Here it is the imposition of external stability which implies that if the SPE σ_p cannot for any history be profitably renegotiated to an SPE admitted by the PP norm, then the norm must admit σ_p; i.e., σ_p is Pareto-perfect. In Example 2, in contrast, the existence of multiple PP norms is essential for showing that the SPE σ_0 is Pareto-perfect. Each of these norms may then be *nonstationary* in the sense of letting the set of SPEa admitted in a subgame depend on the history leading up to that subgame. A PP norm that admits σ_0 for the initial history must have the property that for a history after which a punishment path, say π_1, is to be played, σ_0 can no longer be admitted, else internal stability is violated. The reason why the players respect the norm and let the profitable renegotiation possibility pass is explained by the external stability of the norm: If they were to renegotiate to σ_0, and at some future contingency, a punishment path, say π_2, were to be played, then the players would be able to renegotiate back to σ_0 in a *norm-observing* manner. Hence, given the norm and allowing for renegotiation, σ_0 is not viable when not admitted by the norm.

The paper is organized as follows. Section 2 formally defines the concept of a PP norm for repeated games. Note that since stationarity is not needed, the concept could at a notational cost have been defined for multistage games that are not strictly repeated. Attention is restricted to two-player games since for games with more than two players one would have to consider renegotiation by coalitions short of the grand coalition.[4] Section 3 relates the set of PPEa admitted by *stationary* PP norms to the concepts of F&M and B&R, while Section 4 explores *nonstationary* PP norms by returning to Example 2. Proofs are contained in Section 5.

[4] See Asheim (1988) where Bernheim *et al.*'s (1987) definition of a *Perfectly coalition-proof* equilibrium is extended to infinite horizon multistage games. This generalization is based on Greenberg's (1990) *Theory of Social Situations*.

2. Definition of Pareto-Perfect Norms

The repeated game G consists of a T-fold play of the two-person normal form game $\Gamma := (A_1, A_2, v_1, v_2)$, where T is finite or infinite, and where A_i is the compact mixed[5] action set and v_i the continuous payoff-function of player i, $i = 1, 2$. Writing $A = A_1 \times A_2$, the set of *histories* (subgames) of G is given by $H = \bigcup_{t=0}^{T-1} A^t$. Follow the convention that $A^0 = \{0\}$ and $(0, h) = (h, 0) = h$ for any $h \in H$. The set H is naturally ordered by $\leq$, i.e., $h \leq h'$ means that h equals or precedes h'.

A (strategy) *profile* of G is an element of a set $X = X_1 \times X_2$, where X_i is the set of mappings from H to A_i. A profile $x \in X$ determines a path $\pi(x) = (a^0(x), \ldots, a^{T-1}(x)) \in A^T$ and a payoff for each player i given by $u_i(x) := (1/T) \cdot \sum_{t=0}^{T-1} v_i(a^t(x))$ if $T < \infty$ and $u_i(x) := (1 - \delta) \cdot \sum_{t=0}^{\infty} \delta^t \cdot v_i(a^t(x))$ with $\delta \in (0, 1)$ if $T = \infty$. If $x, y \in X$, y is said to *dominate* x if $u_i(x) < u_i(y)$ for $i = 1, 2$. Denote by E the set of Subgame-perfect equilibria of G and by $U(E) := \{u(x) \mid x \in E\} \subseteq \mathbb{R}^2$ the set of SPE payoffs.

For any $h \in H$, let G^h denote the continuation game of G to be played given h (i.e., a $(T - t)$-fold play of Γ if $h \in A^t$), with X^h and E^h being the corresponding sets of profiles and SPEa. Hence, $G^h = G^{h'}$ if $h, h' \in A^t$, and $G^h = G$, $X^h = X$, and $E^h = E$ for any $h \in H$ if $T = \infty$. If $x \in X$, then $x^h \in X^h$ denotes the continuation profile that x induces on the subgame h (i.e., $x^h(k) = x(h, k)$ for any $k \geq 0$ where $(h, k) \in A^{t+s}$ is a history consisting of $h \in A^t$ followed by $k \in A^s$).

A *social norm* for G is a correspondence Σ assigning to each subgame $h \in H$ a subset, $\Sigma(h)$, of E^h. A norm Σ is said to be *internally stable* against renegotiation if

(IS) For any $h \in H$ and any $x \in \Sigma(h)$, there do not exist $k \geq 0$ and $y \in \Sigma(h, k)$ such that y dominates x^k.

A norm Σ is said to be *externally stable* against renegotiation if

(ES) For any $h \in H$ and any $x \in E^h \backslash \Sigma(h)$, there exist $k \geq 0$ and $y \in \Sigma(h, k)$ such that y dominates x^k.

A social norm is said to be *Pareto-perfect* if it is both internally and externally stable against renegotiation.[6]

For a given repeated game G it is desirable to establish the existence and uniqueness of a PP norm. Existence and uniqueness *can* be estab-

[5] For the general analysis of this paper we allow for private randomizations provided they are publicly observable before the start of the next period. In Examples 1 and 2, however, attention is for simplicity restricted to pure strategies only.

[6] The graph of a PP norm can be shown to be a von Neumann and Morgenstern (vN&M) abstract stable set of a system associated with the game. See Greenberg (1990) for an approach to equilibrium analysis based on stability in the sense of a vN&M abstract stable set.

lished if G is *finitely* repeated. Then the unique PP norm admits SPEa which are *Pareto-perfect*[7] in the sense of Definition 1′ of Section 5. This definition is due to Bernheim and Ray (1985).

PROPOSITION 1. *If G is a finitely repeated game, then there exists a unique Pareto-perfect social norm, Σ, for G. Furthermore, for all $h \in H$, $\Sigma(h)$ is the set of Pareto-perfect equilibria of G^h (according to Definition 1′ of Section 5).*[8]

The definition of a PPE reproduced as Definition 1′ in Section 5 cannot be generalized to infinitely repeated games since it is based on backward recursion from the last stage of the game. The characterization of a PPE given in Proposition 1 is, however, not based on backward recursion. Therefore, this characterization yields a definition of renegotiation-proofness covering both finitely and infinitely repeated games.

DEFINITION 1. Consider a finitely or infinitely repeated game G. An SPE σ is said to be a Pareto-perfect equilibrium of G if and only if there is a Pareto-perfect social norm, Σ, for G such that $\sigma \in \Sigma(0)$.

In infinitely repeated games, there is not, in general, a unique PP norm as illustrated by Example 2. In Definition 1 above, an SPE σ has been defined to be Pareto-perfect if $\sigma \in \Sigma(0)$ for at least one PP norm, Σ. Alternatively, one could require $\sigma \in \Sigma(0)$ for any PP norm, Σ. In Example 2, this latter definition would have lead to the nonexistence of PPEa, while the former admits exitence. *The choice among multiple PP norms should be considered exogenous to the game and hence not open for (re)negotiation by the players.*[9] No theory for selecting among multiple PP norms (i.e., refining the concept of a PPE) is offered.

A norm Σ for an infinitely repeated game G is said to be *stationary* if $\Sigma(h) = \Sigma(0)$ for all $h \in H$. Even though an infinitely repeated game is of a stationary nature, it does not generally follow that a PP norm for such a game is stationary. However, a PP norm can be nonstationary only if there exist multiple PP norms.

[7] This concept is called *Internally consistent* by B&R and *Relative strongly renegotiation-proof* by Farrell and Maskin (1987). It is also contained in van Damme (1987). See Benoit and Krishna (1988) for a slightly different concept.

[8] This characterization of a PPE for finitely repeated games is inspired from Greenberg's (1989) derivation of Bernheim *et al.*'s (1987) *Coalition-proof* equilibrium.

[9] As with DeMarzo's (1988) concept of a *Strong sequential* equilibrium, a norm is influential even after a deviation by the grand coalition. However, here a deviating SPE is nonviable only if at some feasible future contingency the grand coalition *can agree to stop* the deviation and return to the norm. DeMarzo's norms are much more influential in the sense that a deviation is nonviable if the deviating coalition in some stage *cannot agree to* continue the deviation from the norm (i.e., renegotiation of an original profile is not allowed, only coordination in each stage).

PROPOSITION 2. *If for an infinitely repeated game G there exists a unique Pareto-perfect social norm Σ, then Σ is stationary.*

Allowing for nonstationary PP norms ensures existence in Example 2. A general existence result is not available at this point and must await further research.

3. STATIONARY PARETO-PERFECT NORMS

For the definitions below, recall the following preliminaries. Based on Abreu *et al.* (1986) and van Damme (1987), denote by B the *self-enforcement* correspondence for an infinitely repeated game G: $B(V) := \{v \in \mathbb{R}^2 | \exists f\colon A \curvearrowright V \subseteq \mathbb{R}^2$ *such that v is a Nash equilibrium payoff of the game with payoffs* $(1 - \delta)v(a) + \delta f(a)\}$. A set of payoffs V is said to *support* the payoff v if $v \in B(V)$, and furthermore, V is said to be *self-generating* if $V \subseteq B(V)$. If $x \in E$, then $C(x) := \{u(x^h) | h \in H\}$ is bounded and self-generating. Conversely, if V is bounded and self-generating, then for any $v \in V$, there exists some $x \in E$ with $u(x) = v$ and $C(x) \subseteq V$. Say that $x \in E$ is *generated* by V if $C(x) \subseteq V$. Following B&R, say that V *directly dominates* V' (written $V\ d\ V'$) if there exists $v \in V$, $v' \in V'$ such that $v \gg v'$, while V *indirectly dominates* V' if there exists a finite sequence $V_1, \ldots, V_m$ such that $V\ d\ V_1\ d \cdots d\ V_m\ d\ V'$. Denote by V^+ the Pareto-boundary of V: $V^+ = \{v \in V|$ there is no $v' \in V$ such that $v' \gg v\}$.

DEFINITION 2 (F&M, B&R). An SPE σ of an infinitely repeated game G is said to be Weakly renegotiation-proof if and only if there do not exist $h, h' \in H$ such that σ^h dominates $\sigma^{h'}$.

It follows that σ is a WRP equilibrium if and only if σ is generated by some bounded payoff set W satisfying $W^+ = W \subseteq B(W)$. Call a bounded payoff set W WRP if $W^+ = W \subseteq B(W)$. Following Ray (1989), a bounded payoff set I is called *Internally renegotiation-proof* (IRP) if $I = B(I)^+$. Note that any IRP set is WRP (since $I = B(I)^+$ implies $I^+ = I \subseteq B(I)$), while the converse is not true.

DEFINITION 3 (F&M). A WRP equilibrium σ of an infinitely repeated game G is said to be Strongly renegotiation-proof if and only if there do not exist $h \in H$ and a WRP equilibrium x such that x dominates σ^h.

It follows that σ is an SRP equilibrium if and only if σ is generated by some WRP set S which is not directly dominated by any other WRP set. Call a WRP set S SRP if S is not directly dominated by any other WRP set.

DEFINITION 4 (B&R). A WRP set R is said to be Consistent if and only if R indirectly dominates W whenever W is a WRP set indirectly dominating R.

An SPE σ generated by a Consistent set R is called a *Consistent* equilibrium.

If a_n is a Nash equilibrium of the stage game, then $\sigma(\pi_n, \pi_n, \pi_n)$ with $\pi_n = (a_n, a_n, \ldots)$ is WRP. A Consistent equilibrium also exists under general conditions, while there are examples available with no SRP equilibrium (e.g., Example 1 of B&R). However, F&M provide for the latter concept a sufficient condition for existence for large enough δ, a condition that covers a large class of games.

If there exists a *stationary* PP norm, the PPEa admitted by such a norm are related to the concepts above in the following manner.

PROPOSITION 3. *Let* Σ *be a stationary Pareto-perfect social norm for an infinitely repeated game* G. *Then* (a) $U(\Sigma(0))$ *is an IRP set, which implies that* $U(\Sigma(0))$ *is a WRP set,* (b) $S \subseteq U(\Sigma(0))$ *for any SRP set* S, *and* (c) $U(\Sigma(0))$ *directly dominates* W *whenever* W *is a WRP set directly dominating* $U(\Sigma(0))$.

If there exists a *unique* (hence, stationary by Prop. 2) PP norm, the set of PPEa can be characterized as follows (compare with Definitions 2 and 3): *An SPE* σ *of an infinitely repeated game* G *is said to be Pareto-perfect if and only if there do not exist* $h \in H$ *and a Pareto-perfect equilibrium* x *such that* x *dominates* σ^h.

Weak renegotiation-proofness ensures internal stability, but due to its lack of external stability it seems too weak. In particular, van Damme (1989) shows that in a repeated prisoners' dilemma, it does not rule out any SPE payoff. Definitions 1, 3, and 4 represent different ways of imposing external stability. The concepts defined therein coincide in the special case where any payoff efficient within the set of SPE payoffs can be supported by payoffs which are themselves efficient within the set of SPE payoffs.

PROPOSITION 4. *Consider an infinitely repeated game* G *for which* $U(E)^+ \subseteq B(U(E)^+)$. *Then there exists a unique and stationary Pareto-perfect social norm* Σ. *Furthermore,* $\Sigma(0) = \{\sigma \in E | \ C(\sigma) \subseteq U(E)^+\}$. *Finally, the sets of Pareto-perfect, SRP, and Consistent equilibria coincide and equal* $\Sigma(0)$.

Van Damme's (1989) repeated prisoners' dilemma satisfies the above condition.

As demonstrated by Example 1, illustrated in Fig. 1, it is not a general property that the concepts of Pareto-perfect, SRP, and Consistent equilibria coincide.

Example 1. This game consists of the infinite repetition of

	a_{2m}	a_{2n}	a_{2p}	a_{2q}
a_{1m}	1, 3	−8, −8	−8, 0	−8, −8
a_{1n}	0, −8	2, 4	4, 0	7, −8
a_{1p}	−8, −8	−8, −8	3, 1	−8, −8
a_{1q}	−8, −8	−8, −8	−8, 0	4, 2

with discount factor $\delta = \frac{1}{2}$, and where mixed strategies are *not* admitted.

Write $a_j = (a_{1j}, a_{2j})$ and $\pi_j = (a_j, a_j, \ldots), j = m, n, p, q$. Using Definitions 3 and 4 one can now show that $\sigma_n = \sigma(\pi_n, \pi_n, \pi_n)$ is the unique SRP and Consistent equilibrium. Still, the unique and stationary PP norm also admits $\sigma_p = \sigma(\pi_p, \pi_n, \pi_n)$.

Claim 1. *For the game of Example* 1 *there is a unique and stationary Pareto-perfect norm* Σ. *Furthermore*, $\Sigma(0) = \{\sigma \in E | \sigma(h) \in \{a_n, a_p\}$ *for any* $h \in H\}$.

By Proposition 3(b) it is general property that Strong renegotiation-proofness is stronger than Pareto-perfectness (according to Def. 1), provided that a stationary PP norm exists. One can argue that the requirement of Strong renegotiation-proofness is too strong for the same reason that Rubinstein's (1980) concept of a *Strong perfect* equilibrium is too strong. In a two-player game, $\sigma \in E$ is a Strong perfect equilibrium if there do not exist $h \in H$ and $x \in X$ such that x dominates σ^h. This concept is too strong because, when not all profiles are considered viable, the existence of any profile x that dominates σ^h cannot imply that σ^h would be renegotiated if h were to be reached. It follows that we do not expect that allowing for renegotiation necessarily will take us to profiles that are not dominated by any profile. In the same vein, when not all WRP equilibria are considered viable, then the existence of any WRP equilibrium x that dominates σ^h cannot imply that σ^h would be renegotiated if h were to be reached. As above, we do not expect that allowing for renegotiation necessarily will take us to WRP equilibria that are not dominated by any WRP equilibrium.

Example 1 illustrates this point. Here the players do not wish to renegotiate σ_p in favor of the nonviable WRP equilibrium $\sigma_q = \sigma(\pi_q, \pi_m, \pi_m)$. Hence, σ_p should be considered renegotiation-proof even though it is dominated by some WRP equilibrium.

By permitting indirect domination (compare Def. 4 with Prop. 3(c)), the concept of a Consistent equilibrium also uses σ_q to eliminate σ_p: $C(\sigma_n)$ indirectly dominates $C(\sigma_p)$ (through $C(\sigma_q)$) while the converse is not true.

Again one can argue that a nonviable WRP equilibrium like σ_q should not be used to upset σ_p.[10]

4. Nonstationary Pareto-Perfect Norms

Requiring Weak renegotiation-proofness entails an assumption of *stationarity* in the sense that the viability of an SPE x for the initial history is made to imply the viability of x in every subgame. By again considering Example 2, illustrated in Fig. 2, the inappropriateness of such an assumption is brought forward.

Example 2. This game consists of the infinite repetition of

	a_{20}	a_{21}	a_{22}	a_{23}
a_{10}	3, 3	−5, −5	−5, −5	−5, 4
a_{11}	−5, −5	1, 2	−5, −5	−5, 3
a_{12}	−5, −5	−5, −5	2, 1	−5, 2
a_{13}	4, −5	2, −5	3, −5	0, 0

with discount factor $\delta = \frac{1}{2}$, and where mixed strategies are *not* admitted.

Write $a_j = (a_{1j}, a_{2j})$ and define π_j and $\sigma_j, j = 0, 1, 2, 3$, as in the Introduction. Using Definition 2 one can now show that σ_3 with payoffs (0, 0) is the unique WRP equilibrium; hence, it is also the unique SRP equilibrium as well as the unique Consistent equilibrium. However, σ_0, σ_1, and σ_2 are three SPEa each of which, for every history, dominates the unique SPE admitted by the concepts of F&M and B&R. We proceed to make the following claim.

Claim 2. *In the game of Example* 2, σ_0, σ_1, *and* σ_2 *are all Pareto-perfect equilibria, while* σ_3 *is not.*

The proof of Claim 2 shows that σ_j, $j = 0, 1, 2$, correspond to three separate *nonstationary* PP norms, each of which admits only σ_j^h in the subgame h.

[10] B&R motivate their definition by a three-player game (their Example 2) in which the relevant WRP sets dominate each other cyclically: $W_1\ d\ W_3\ d\ W_2\ d\ W_1$. Hence, if the Consistence of a WRP set R required that R directly dominates W whenever W is a WRP set directly dominating R, then no nonempty Consistent set would exist. By instead requiring only *in*direct domination, $W_j, j = 1, 2, 3$, all become Consistent. Even though the concept of Pareto-perfectness (according to Def. 1) does not permit such indirect domination, by allowing for nonstationary PP norms it can, in this example, be shown that for any $v \in W_1 \cup W_2 \cup W_3$, there exists a PPE σ such that $u(\sigma) = v$. (Proof of this claim is available on request from the author.)

Assuming the adoption of the PP norm corresponding to $\sigma_0 = \sigma(\pi_0, \pi_1, \pi_2)$, the following informal story explains why attempts to renegotiate after a deviation from π_0 by, say, player 1 must fail given this particular norm. After such a deviation, the norm uniquely prescribes following π_1. The grand coalition profits if it is able to renegotiate to an SPE yielding π_0. Such an alternative SPE must, however, in turn be supported by a threat of *not* continuing π_0 if say, player 2 immediately deviates to (a_{10}, a_{23}). Since (a_{10}, a_{23}) is a *simultaneous* deviation from the first-period play of π_1 (i.e., a_1) and hence is not punished by the simple strategy profile $\sigma(\pi_0, \pi_1, \pi_2)$, the adopted norm at such a contingency uniquely prescribes the second-period continuation of π_1 (i.e., π_0). Hence, if player 2, after the initial deviation from π_0 by player 1, suggests renegotiating to an SPE yielding π_0, player 1 will not go along, realizing that player 2 gains by immediately deviating to (a_{10}, a_{23})—reaping a payoff in that stage of four instead of three (at a severe cost to player 1)—then securing through mutually advantageous and norm-observing renegotiation that π_0 be followed thereafter.[11]

Claim 2 establishes the following proposition.

PROPOSITION 5. *Consider an infinitely repeated game G. Then an SPE x being a WRP/SRP/Consistent equilibrium is neither sufficient nor necessary for x being a Pareto-perfect equilibrium according to Definition* 1.

The problem with the concepts of F&M and B&R in the context of Example 2 is that being a WRP equilibrium is considered a prerequisite for viability, such that nonviable SPEa need not be defeated through renegotiation by viable ones.

In Example 2, our results are in the spirit of Pearce's (1987) approach to renegotiation-proofness which has been further developed by Abreu *et al.* (1989) (APS) for the case of symmetric games:

DEFINITION 5. (APS). An SPE σ of a symmetric infinitely repeated game G is said to be a Consistent bargaining equilibrium (CBE) if and only if $l(x) \leqq l(\sigma)$ for any SPE x, where the function l is given by $l(x) := \inf\{\min\{v_1, v_2\} | (v_1, v_2) \in C(x)\}$.

Hence, by this definition, the punishments that support a CBE are not renegotiated because the players realize that any SPE must involve at least as hard a punishment for some player in some subgame. Example 2 shows a symmetric game (interchange the two middle rows) for which $l(x) \leqq 2$ for any $x \in E$. Hence, the PPEa σ_0, σ_1, and σ_2 are all CBE since $l(\sigma_0) = l(\sigma_1) = l(\sigma_2) = 2$. This indicates that the concept of a CBE may be given an interpretation in terms of nonstationarity.[12]

[11] This verbal explanation was suggested to me by Joseph Farrell.

[12] See Bergin and MacLeod (1989) for the role of *stationarity* in an axiomatic approach to renegotiation-proofness. Note that the comparison with the concept of a CBE in Example 2

Example 2 illustrates a case in which it is hard to impose renegotiation-proofness since any SPE but the unanimously least attractive has continuation equilibria which dominate each other. Due to this inherent difficulty, the concept of a PPE has little bite in refining the concept of a SPE (see footnote 12). This contrasts Example 1 as well as the class of games covered by Proposition 4, which due to their uncomplicated nature give rise to a unique PP norm yielding a substantial refinement of subgame-perfectness.

5. Proofs

Lemma 1. *Let G be a finitely or infinitely repeated game. If Σ is a Pareto-perfect social norm for G, then $\sigma \in \Sigma(0)$ implies $\sigma^h \in \Sigma(h)$ for any $h \in H$.*

Proof. Let $x \in E$. Suppose $x^h \notin \Sigma(h)$ for some $h \in H$. By (ES), there exist $h' \geq h$ and $\sigma \in \Sigma(h')$ such that σ dominates $x^{h'}$. By (IS), $x \notin \Sigma(0)$. ■

For the proof of Proposition 1 we need to reproduce the following definition of a *Pareto-perfect* equilibrium of a finitely repeated game G.

Definition 1′ (Bernheim and Ray, 1985; B&R). (i) Let $h \in A^{T-1}$. Then $\sigma \in E^h$ is a PPE of G^h if and only if there is no $x \in E^h$ such that x dominates σ.

(ii) Let $h \in A^t$ with $0 \leqq t < T - 1$. Assume that PPE has been defined for all games $G^{h'}$ with $h' > h$. Then $\sigma \in E^h$ is a PPE of G^h if and only if

(a) σ^k is a PPE of $G^{(h,k)}$ for all $k > 0$, and
(b) there is no $x \in E^h$ satisfying part (a) such that x dominates σ.

Let P^h denote the set of PPEa of G^h (according to Definition 1′). Also let $Q^h := \{\sigma \in E^h | \sigma^k \in P^{(h,k)} \text{ for all } k > 0\}$. By adapting the inductive proof of van Damme's (1987) Theorem 8.7.1, it follows that $U(Q^h)$ is compact.

Lemma 2. *For any $h \in H$ and any $\sigma \in P^h$, there do not exist $k \geq 0$ and $x \in P^{(h,k)}$ such that x dominates σ^k.*

Proof. If $\sigma \in P^h$ and $k \geq 0$, then $\sigma^k \in P^{(h,k)}$ by Definition 1′(ii)(a). That $x \in P^{(h,k)}$ and x dominates σ^k is impossible by Definition 1′(i) and (ii)(b). ■

should not be drawn too far: Although σ_3 with payoffs (0, 0) is *not a* PPE, there is a PPE σ' with payoffs (0, 0), being described by: (i) start with π_3, and (ii) reward (!) any individual deviation from π_3 with the continuation equilibrium σ_0. The fact that the SPE σ', which is *not* a CBE, corresponds to a nonstationary PP norm is shown in the same way as in the proof of Claim 2.

LEMMA 3. *For any $h \in H$ and any $x \in E^h \backslash P^h$, there exist $k \geq 0$ and $\sigma \in P^{(h,k)}$ such that σ dominates x^k.*

Proof. If $x \in E^h \backslash P^h$, then, by Definition 1′, there exist $k \geq 0$ and $\sigma \in Q^{(h,k)}$ such that σ dominates x^k. Definition 1′(i) and (ii)(b) implies that, since $U(Q^{(h,k)})$ is compact, we may w.l.o.g. choose $\sigma \in P^{(h,k)}$. ■

Let Σ defined by $\Sigma(h) = P^h$ for all $h \in H$ be a social norm for G.

COROLLARY. *Σ is a Pareto-perfect social norm for G.*

Proof. Lemmas 2 and 3 are identical to (IS) and (ES). ■

Proof of Proposition 1. By the corollary it remains to be shown that the PP norm Σ is unique. Therefore, let $\tilde{\Sigma}$ be any PP norm for G. Let $h \in H$ satisfy $\tilde{\Sigma}(h') = P^{h'} = \Sigma(h')$ for all $h' > h$. Definition 1′ implies that for any $\sigma \in P^h$, there is no $x \in Q^h$ dominating σ. Thus, by Lemmas 1 and 2, for any $\sigma \in P^h$, there do not exist $k \geq 0$ and $y \in \tilde{\Sigma}(h, k) \subseteq Q^{(h,k)}$ such that y dominates σ^k; i.e., $\sigma \in \tilde{\Sigma}(h)$ by (ES) of $\tilde{\Sigma}$. Then, by Lemma 3, for any $x \in E^h \backslash P^h$, there exist $k \geq 0$ and $y \in P^{(h,k)} \subseteq \tilde{\Sigma}(h, k)$ such that y dominates x^k; i.e., $x \notin \tilde{\Sigma}(h)$ by (IS) of $\tilde{\Sigma}$. Hence, $\tilde{\Sigma}(h) = P^h = \Sigma(h)$. By induction from the last stage of G, it follows that $\tilde{\Sigma} = \Sigma$. ■

Proof of Proposition 2. Suppose Σ is PP, but not stationary. Then there exist $h > 0$ such that $\Sigma(h) \neq \Sigma(0)$. Define $\tilde{\Sigma}$ by the property that $\tilde{\Sigma}(k) = \Sigma(h, k)$ for any $k \geq 0$. It follows from the stationarity of G that $\tilde{\Sigma}$ ($\neq \Sigma$) is PP. ■

Proof of Propostion 3. Write $P := \Sigma(0)$ ($= \Sigma(h)$ for all $h \in H$).
(*a*) *By Lemma* 1,

$$U(P) \subseteq B(U(P)), \tag{1}$$

and furthermore, for any $v \in B(U(P))^+$, there exists some $\sigma \in E$ with $u(\sigma) = v$ and $\sigma^h \in P$ for all $h > 0$. By (IS), there do not exist $h > 0$ and $x \in P$ such that x dominates σ^h. Thus, by (1) and (ES), $\sigma \in P$; i.e., $B(U(P))^+ \subseteq U(P)$. We need to establish that $U(P) \subseteq B(U(P))^+$. By (1), it is sufficient to show that

$$U(P) \cap (B(U(P)) \backslash B(U(P))^+) = \emptyset. \tag{2}$$

Therefore, suppose there exists $x' \in P$ with $u(x') \in B(U(P)) \backslash B(U(P))^+$. Then

(i) by (IS), there does not exist $y' \in P$ such that y' dominates x', and

(ii) by Lemma 1, there exists $y \in E$ with $y^h \in P$ for all $h > 0$ such that y dominates x'.

By (IS), there do not exist $h > 0$ and $x \in P$ such that x dominates y^h. By (i), there does not exist $y' \in P$ such that y' dominates y. Thus, by (ES), $y \in P$. Since y dominates x', this contradicts (i). Hence, no such x' exists and (2) follows.

(b) By (a), $U(P)$ is a WRP set, and hence, does not directly dominate the SRP set S. It follows from (ES) that $\sigma \in P$ for any σ generated by S.

(c) Suppose $U(P)$ does not directly dominate W. It follows from (ES) that $\sigma \in P$ for any σ generated by W. Since $U(P)^+ = U(P)$, W does not directly dominate $U(P)$. ■

Proof of Proposition 4. Let $\tilde{\Sigma}$ be any PP norm. If $\sigma \in \Sigma(h)$, there do not exist $k \geq 0$ and $x \in E \supseteq \tilde{\Sigma}(h, k)$ such that x dominates σ^k. This establishes (IS) of Σ and shows by (ES) of $\tilde{\Sigma}$ that $\Sigma(h) \subseteq \tilde{\Sigma}(h)$ for any $h \in H$. For the class of games considered, $U(E)$ is nonempty and compact (van Damme, 1987, Thm. 8.5.6). Hence, by the premise, if $x \in E\backslash\Sigma(h)$, there exist $k \geq 0$ and $\sigma \in \Sigma(h, k) \subseteq \tilde{\Sigma}(h, k)$ such that σ dominates x^k. This establishes (ES) of Σ and shows by (IS) of $\tilde{\Sigma}$ that $\tilde{\Sigma}(h) \subseteq \Sigma(h)$ for any $h \in H$. The sets of SRP and Consistent equilibria equal $\Sigma(0)$ since there exists no WRP set directly dominating the WRP set $U(E)^+$, while if W is a WRP set such that $W \nsubseteq U(E)^+$, then $U(E)^+$ directly dominates W. ■

Proof of Claim 1. Note first that there can be no SPE play off the diagonal. Also, Σ satisfies (IS). Let $\tilde{\Sigma}$ be any PP norm. By ES of $\tilde{\Sigma}$, $\sigma_n \in \tilde{\Sigma}(h)$, since there do not exist $k \geq 0$ and $x \in E \supseteq \tilde{\Sigma}(h, k)$ such that x dominates σ_n^k. By (IS) of $\tilde{\Sigma}$, $x \notin \tilde{\Sigma}(h)$ if there exists k' such that $x(k') = a_q$ since punishing player 1 requires that for some $k \geq 0$, $\sigma_n \in \tilde{\Sigma}(h, k)$ dominates x^k. This shows by (ES) of $\tilde{\Sigma}$ that $\Sigma(h) \subseteq \tilde{\Sigma}(h)$ for any $h \in H$. Finally, if there exists k' such that $x(k') = a_m$, then there exist $k \geq 0$ and $\sigma \in \Sigma(h, k) \subseteq \tilde{\Sigma}(h, k)$ such that σ dominates x^k. This establishes (ES) of Σ and shows by (IS) of $\tilde{\Sigma}$ that $\tilde{\Sigma}(h) \subseteq \Sigma(h)$ for any $h \in H$. ■

Proof of Claim 2. Note first that there cen be no SPE play off the diagonal. Let Σ_j, $j = 0, 1, 2$, be defined by: for any $h \in H$, $\Sigma_j(h) = \{\sigma_j^h\} = \{(\sigma(\pi_j, \pi_1, \pi_2))^h\}$. (IS) follows since, for any $h \in H$, $\Sigma_j(h)$ contains a single element, viz., σ_j^h. To show (ES), suppose x satisfies: there does not exist $h \in H$ such that σ_j^h dominates x^h. (i) For h such that $\sigma_j(h) = a_0$, $x(h) = a_0$, since otherwise σ_j^h dominates x^h. (ii) For h such that $\sigma_j(h) = a_1$, $x(h) \in \{a_0, a_1, a_2\}$. But $x(h) = a_0$ is not possible since $h' = (h, (a_{10}, a_{23}))$ and x being an SPE would imply $\pi(x^{h'}) \neq \pi^0$. However, $\pi(\sigma_j^{h'}) = \pi_0$ since the simultaneous deviation to (a_{10}, a_{23}) is not punished, implying that $\sigma_j^{h'}$ dominates $x^{h'}$. By the same argument, $x(h) \neq a_2$. Hence, $x(h) = \sigma_j(h)$. (iii) For h such that $\sigma_j(h) = a_2$, we repeat the argument under (ii) to show that $x(h) = \sigma_j(h)$. Hence, $x = \sigma_j$. Similarly, $x = \sigma_j^h$ if x satsifies: there does not exist $k \geq 0$ such that $\sigma_j^{(h,k)}$ dominates x^k. This shows (ES). Thus,

Σ_j, $i = 0, 1, 2$, are nonstationary PP norms and $\sigma_j, j = 0, 1, 2$, are PPEa according to Definition 1.

In order to show that $\sigma_3 = \sigma(\pi_3, \pi_3, \pi_3)$ is not a PPE, suppose that there exists some PP norm, Σ_3, such that $\sigma_3 = \sigma_3^h \in \Sigma_3(h)$ for all $h \in H$ (see Lemma 1). Then, *either* there is an SPE $x \in \Sigma_3(h)$ generating strictly positive payoffs, which contradicts (IS) of Σ_3, *or* there is no such SPE, which contradicts (ES) of Σ_3. ■

References

Abreu, D. (1988). "On the Theory of Infinitely Repeated Games with Discounting," *Econometrica* **56,** 383–396.

Abreu, D., Pearce, D., and Stacchetti, E. (1986). "Optimal Cartel Equilibria with Imperfect Monitoring," *J. Econ. Theory* **39,** 251–269.

Abreu, D., Pearce, D., and Stacchetti, E. (1989). "Renegotiation and Symmetry in Repeated Games," ST/ICERD Discussion Paper TE/89/198, London School of Economics.

Asheim, G. B. (1988). "Renegotiation-Proofness in Finite and Infinite Stage Games through the Theory of Social Situations," Discussion Paper A-173, University of Bonn.

Benoit, J.-P., and Krishna, V. (1988). "Renegotiation in Finitely Repeated Games," Working Paper 89-004, Harvard Business School.

Bergin, J., and MacLeod, B. (1989). "Efficiency and Renegotiation in Repeated Games," Discussion Paper 752, Queen's University.

Bernheim, B. D., Peleg, B., and Whinston, M. D. (1987). "Coalition-Proof Equilibria. I. Concepts," *J. Econ. Theory* **42,** 1–12.

Bernheim, B. D., and Ray, D. (1985). "Pareto Perfect Nash Equilibria," mimeo, Stanford University.

Bernheim, B. D., and Ray, D. (1989). "Collective Dynamic Consistency in Repeated Games," *Games Econ. Behav.* **1,** 295–326.

DeMarzo, P. M. (1988). "Coalitions and Sustainable Social Norms in Repeated Games," IMSSS Technical Report 529, Stanford University.

Farrell, J., and Maskin, E. (1987). "Renegotiation in Repeated Games," Working Paper 8759, Department of Economics, University of California, Berkeley.

Farrell, J., and Maskin, E. (1989). "Renegotiation in Repeated Games," *Games Econ. Behav.* **1,** 327–360.

Greenberg, J. (1989). "Deriving Strong and Coalition-Proof Nash Equilibria from an Abstract System," *J. Econ. Theory* **49,** 195–202.

Greenberg., J. (1990). *The Theory of Social Situations: An Alternative Game-Theoretic Approach.* Cambridge: Cambridge Univ. Press.

Pearce, D. (1987). "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," Cowles Foundation Discussion Paper No. 855, Yale University.

Ray, D. (1989). "Internally Renegotiation-Proof Equilibrium Sets: Limit Behaviour with Low Discounting," mimeo, Indian Statistical Institute.

Rubinstein, A. (1980). "Strong Perfect Equilibrium in Supergames," *Int. J. Game Theory* **9,** 1–12.

Selten, R. (1973). "A Simple Model of Imperfect Competition Where 4 Are Few and 6 Are Many," *Int. J. Game Theory* **3,** 141–201.

van Damme, E. E. C. (1987). *Stability and Perfection of Nash Equilibria.* Berlin: Springer-Verlag.

van Damme, E. E. C. (1989). "Renegotiation-Proof Equilibria in Repeated Prisoners' Dilemma," *J. Econ. Theory* **47,** 206–217.