UNIVERSITY OF CALIFORNIA AT BERKELEY

Department of Economics

Berkeley, California 94720

Working Paper No. 91-163

Reneging and Renegotiation

Matthew Rabin

Economics Department

University of California at Berkeley

April 1991

Key words: Renegotiation, repeated games, credibility

Abstract

I present a model of renegotiation in infinitely repeated games that incorporates the assumption that if a player cheats early in a game, then other players are likely to distrust his promises not to cheat in the future. I show how this can allow much more cooperation in some games than allowable in the approach taken by Bernheim and Ray [1989] and Farrell and Maskin [1989]. In other games, my solution concept rules out *all* Pareto-efficient subgame-perfect equilibria, even as the discount factor approaches 1. This contrasts (generically) with existing models of renegotiation.

JEL Classification: 026

I thank Eddie Dekel and Joe Farrell for helpful comments.

Reneging and Renegotiation[0]

I. Introduction

Game theorists frequently assume that players who can negotiate before a game begins will agree to an equilibrium that is Pareto-efficient among the set of equilibria. Recently, the literature on renegotiation has challenged this simple formulation when applied to multi-stage games. It is argued that if players can negotiate efficiently at the beginning of a game, then they likely can do so later in the game. A problem arises: Pareto-efficient equilibria often involve inefficient equilibria in some continuation games. Starting from this problem, the renegotiation literature considers how to combine the idea of efficient negotiations with some notion of time-consistency.

While the general critique is compelling, I believe that approaches in the current renegotiation literature omit an important issue of credibility. In this paper, I define an alternative solution concept for infinitely repeated games that captures this credibility issue.

There are currently two major strands of renegotiation concepts in infinitely repeated games. The first is that of Bernheim and Ray [1989] and Farrell and Maskin [1989], which I shall throughout refer to as the "BRFM approach." Roughly, these papers take the view that--because each continuation game is identical--the set of feasible equilibria ought be identical for each

[0] I thank Eddie Dekel and Joe Farrell for helpful comments.

continuation game of an infinitely repeated game. While behavior may be dependent on the history of play in a game, combining the stationarity of the set of feasible equilibria with the presumption of efficient negotiations implies that no continuation equilibrium for one history can Pareto-dominate the continuation equilibrium after another a history. These papers then elaborate and strengthen these concepts.

A second strand involves those papers building from Pearce [1987], Abreu and Pearce [1989], and Abreu, Pearce, and Stachetti [1989]. In this view, the set of plausible continuation equilibria can be dependent on the previous play of a game. Roughly, if after some history, players attempt to renegotiate away from a particular continuation equilibrium, they are constrained to do so to those equilibria for which all continuation equilibria are themselves superior to the continuation equilibrium being negotiated away from. The motivation for this restriction is fairly compelling: players will tend to renegotiate--as in the BRFM approach--but it isn't "credible" to propose an equilibrium which seems just as susceptible to post-deviation renegotiation as the one players are currently renegotiating away from.

While my approach shares the history dependence feature with the Pearce approach, it does so in a complementary way, capturing a different issue of credibility.

Consider Example 1. This game is similar to the prisoners' dilemma, but contains a pure-strategy equilibrium which is even worse than the "defection equilibrium". If this game is infinitely repeated with a high enough discount factor, a natural outcome to consider is where the two players cooperate each period by playing (U,L).[1] Yet either player could increase his stage-game

[1] While I concentrate in this paper exclusively on infinitely repeated games, RPE and SRPE can apply *without* modification to finitely repeated games

payoff by deviating from playing (U,L). Because cheating is tempting, this cooperation must be enforced by "punishing" in future periods any player who deviates from the cooperative mode.

Example 1

		Player 2		
		L	C	R
	U	4,4	0,5	0,0
Player 1	M	5,0	2,2	0,0
	D	0,0	0,0	1,1

One such punishment is for both players to play (D,R) forever following a deviation. Indeed, if the discount factor is greater than 1/4 but less than 1/3, the only subgame-perfect equilibria yielding the cooperative outcome involve at least one period of playing (D,R) or worse after deviations.

The renegotiation literature raises the following question: If one of the players actually does deviate, will the players really play (D,R)? Would not the players negotiate to a better equilibrium? For instance, they might agree to play the somewhat more favorable (M,C) each period. Both the Pearce approach and the BRFM approach eliminate this equilibrium, and this argument seems to me compelling. After a deviation, players who can readily negotiate

(and even to serial play of different games). For instance, the reader can verify that the arguments in the text apply to Example 1 if we consider finite repetition.

While Pearce [1987] and Farrell and Maskin [1989] also focus exclusively on infitely repeated games, other papers such as Bernheim and Ray [1989] also consider finitely repeated games. As per the arguments in the text, I believe the approach of Bernheim and Ray [1989] exaggerates the problem with renegotiation in finite repetitions of Example 1.

would likely abandon their original punishment, and maybe play (M,C) rather than (D,R). As a consequence, for a discount factor between 1/4 and 1/3, the cooperative equilibrium seems implausible.

Yet the BRFM approach makes a stronger claim: the threat to play (M,C) after a deviation is also considered an implausible way to enforce cooperation. It is argued that if the players can negotiate to full cooperation at the beginning of the game, they could then renegotiate to full cooperation after a deviation.

There is, however, a fundamental difference from the previous case that is obscured if we label *all* inefficient continuation equilibria as punishments. Whereas playing (D,R) forever might reasonably be called a punishment, I do not think playing (M,C) should necessarily be perceived as a punishment. Rather, it could reflect a loss of credibility by one or more of the players.

If you cheat, you cannot renegotiate as if you were negotiating for the first time. If player 1 appears initially to be trustworthy, and promises to be cooperative, player 2 may believe him; cooperation may be plausible. But if player 1 appears initially to be trustworthy, promises to be cooperative, then cheats on this agreement, and then promises to be cooperative, player 2 will likely not believe him. The players cannot renegotiate back to cooperation because player 2 no longer trusts player 1, not because he wants to punish player 1 for deviating. Renegotiation from (M,C) to (U,L) might not occur if player 1 has lost his credibility, even if player 2 does not seek to "punish" player 1.

We can still rule out a punishment of (D,R) because player 2 should agree to renegotiate from (D,R) to (M,C) even though player 1 has lost credibility. Because (M,C) is a strict Nash equilibrium in the stage game, player 2 does not need to rely on trusting player 1. Refusing to renegotiate from (D,R) to

(M,C) would appear to be a time-inconsistent punishment; refusing to renegotiate from (M,C) to (U,L) would seem to be prudent skepticism. Because it ignores how the set of feasible continuation equilibria can depend on the history-contingent credibility of players, the BRFM approach misses the important distinction in these two cases.

In the next section, I develop reneging-proof equilibrium (RPE), which restricts the set of equilibria to those in which cheating players lose credibility, but without any presumption of efficient negotiations. (The name is meant to emphasize the darker side of renegotiation: before the question of renegotiation is likely to arise, some player must renege on a previous agreement.) In Section III, I develop super reneging-proof equilibrium (SRPE) by combining RPE with the presumption of dynamically consistent efficient negotiations.

I do not claim that SRPE is manifestly a good solution concept to apply to games, partly because of its stark formulation of the credibility issue, but mostly because it incorporates the strong presumption of efficient negotiations. The idea that communication (or anything else) leads generally to efficient equilibria has not been strongly supported with research. SRPE incorporates the "presumption of efficiency" solely to illustrate an alternative approach to the renegotiation issues. I compare my approach to others in the renegotiation literature in Section IV. In Section V, I illustrate that SRPE achieves a possibility to which the renegotiation literature alludes: Assuming that all renegotiations will be Pareto-efficient can mean that no equilibrium will be Pareto-efficient, even as the discount factor approaches 1. I conclude the paper in Section VI.

II. Temptation, Cheating, and Credibility

I consider N-player, complete-information games consisting of an infinite repetition of a stage game G. I assume that G has a finite strategy space $A = A_1 \times \cdots \times A_N$, where A_i is the set of mixed strategies for player i. Let the function $u_k(a)$ represent player k's stage-game payoffs from strategy profile $a \in A$. From the stage-game payoffs, we can construct the payoffs from any play of the game: Letting $a = \{a^1, a^2, \ldots, a^t, \ldots\} \in A^\infty$ be some path of play, where a^t is the play of the game in period t, the payoff to player k will then be $U_k(a) = (1-\delta)\cdot\sum_t \delta^t \cdot u_k(a^t)$, where $\delta < 1$ is the discount factor. I shall refer to the repeated game with stage game G and discount rate δ as $\Gamma(G,\delta)$.

Consider a strategy profile $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ for $\Gamma(G,\delta)$. The payoffs to a strategy profile σ of $\Gamma(G,\delta)$ is given by $v_k(\sigma) \equiv \sum_{a \in A^\infty} \sigma_a U_k(a)$, where σ_a is the probability induced by σ of the path a.

For any T-period history of play, let $b_T \equiv [b^1, b^2, \ldots, b^T] \in A^T$, where b^t is the mixed-strategy profile in period t. I shall assume throughout that strategies--not only actions--are observable to other players. Let $\sigma^t(b_{t-1})$ be the period t strategy profile implied by the path b_{t-1}, and let $\sigma[b_{t-1}]$ be the continuation strategy on the continuation game following the path b_{t-1}.

Let $\mathcal{E}$ be the set of subgame-perfect equilibria (SPE) in $\Gamma(G,\delta)$. Consider a SPE σ of $\Gamma(G,\delta)$. If in some stage game σ does not induce a Nash equilibrium in that stage game, then some player could gain in that stage game by changing his strategy. Thus, the player has the <u>temptation</u> to cheat: this temptation will play a central role in my theory. The temptation to a player in an equilibrium is the biggest short-term gains available from cheating on the equilibrium (and thus the temptation is always non-negative).

Definition 1:

For a SPE, σ, the temptation is $\tau(\sigma) = (\tau_1(\sigma), \tau_2(\sigma), \ldots, \tau_N(\sigma))$, where $\tau_k(\sigma) = \mathrm{Max}_{t, b_{t-1}, a \in A_k^t} \{u_k(a, \sigma_{-k}^t(b)) - u_k(\sigma^t(b))\}$,

where $\sigma_{-k}^t(b)$ is the period t strategy profile given by the equilibrium σ for all the players except k.

The solution concepts I develop build on the idea that the credibility a player has in persuading others that he will resist future temptation may be a function of whether he has resisted previous temptations to cheat. While I shall mainly be interested in how this idea combines with the presumption of efficiency, in this section I define an equilibrium concept which does not incorporate the presumption of efficiency. In a sense, I am first refining SPE to capture the loss of credibility as players deviate from equilibria. Given this modification, I will then in Section III apply "dynamic efficiency" arguments to this modification of SPE.

The first step is characterize the types of deviations from equilibria:

Definition 2:

Consider a SPE σ and history b_t.

Then $\Delta_k(b_t, \sigma) = u_k(b_k^t, \sigma_{-k}^t(b_{t-1})) - u_k(\sigma^t(b_{t-1}))$.

Loosely, $\Delta_k(b_t, \sigma)$ refers to player k's one-shot gain from a given deviation in period t. When this value is positive after a given history of play, player k has not resisted the temptation to cheat. If player k does not cheat in period t--and follows the prescribed equilibrium strategy--the value $\Delta_k(b_t, \sigma)$ equals zero. Note that $\Delta_k(b_t, \sigma)$ can be negative; a "deviation" from an equilibrium can be costly for a player, even in the short term. As will be incorporated in my definitions, such a "deviation" is perhaps likely to be interpreted by other players as a slip, rather than as an intentional deviation, and is not likely to involve a loss of credibility for that player.[2] I shall generally refer to a deviation for which $\Delta_k(b_t,\sigma)$ is positive as cheating, to distinguish it from "slips."

Definition 3 formalizes the idea that a player's credibility depends on his history of cheating.

Definition 3:

Consider a SPE σ and history b_T.

$$\text{Then } C_k(b_T,\sigma) = \begin{cases} 1 \text{ iff for all } t \le T,\ \Delta_k(b_t,\sigma) \le 0. \\ 0 \text{ iff there exists } t \le T \text{ such that } \Delta_k(b_t,\sigma) > 0. \end{cases}$$

We can refer to $C_k(b_t,\sigma)$ as the credibility of player k after history b_t with respect to equilibrium σ. Credibility is defined to take on only two values; if player k has not previously cheated on the equilibrium, he has credibility 1. If he has cheated at any point, his credibility is zero.

The first solution concept incorporates this notion of credibility. An equilibrium is reneging-proof if after any cheating by a player, the continuation equilibrium is not tempting for that player; once a player cheats--reneging on previous promises--he is not trusted any more, and the other players will not rely on him to resist the temptation to cheat again.

[2] Of course, players could be concerned with each other's competence; slips may imply future slips not because it reveals a player as a cheater, but because it reveals him as incompetent.

Definition 4:

A SPE σ is a Reneging-Proof Equilibrium (RPE) iff for all k, for all t, for all histories b_t such that $C_k(b_t,\sigma) = 0$, $\tau_k(\sigma[b_t]) = 0$.

It would seem reasonable to suppose that RPE does not much restrict outcomes compared to SPE: intuitively, losing one's credibility is bad, and seems likely to imply that the "punishments" for cheating are likely to be more rather than less severe if players are assumed to lose credibility. If we do not eliminate any of the most severe punishments, we will not eliminate any possible equilibria.

It turns out, however, that RPE can limit the severity of punishments, and thus rule out some outcomes. The reason is that punishing a player may be sustainable only because the players doing the punishing are worried that if *they* deviate from the punishment, that *they* will be punished. Sometimes, however, the punishment for such "secondary cheating" involves a continuation equilibrium that places trust in the original cheater not to cheat again. By assumption, however, the RPE cannot involve relying on the original cheater to resist any temptation.

Consider Example 2, which has only one Nash equilibrium--(A_2,B_2), yielding payoffs of (10,10).

Example 2

		Player 2			
		B_1	B_2	B_3	B_4
	A_1	11,11	3,15	0,0	0,0
Player 1	A_2	15,3	10,10	15,0	0,0
	A_3	0,0	0,15	0,0	3,8
	A_4	0,0	0,0	8,3	0,0

Suppose this game were repeated infinitely often, with a discount factor of 3/4. One SPE outcome in the repeated game is the cooperative one in which the players play (A_1,B_1) each period. The only way to enforce this equilibrium, however, is to punish a cheater quite severely; playing (A_2,B_2) forever after somebody deviates would not, for instance, be a sufficient deterrent.

Suppose that player 1 has already deviated. Then, in any RPE, the continuation equilibrium cannot contain any period in which player 1 would be tempted to cheat. The only continuation equilibrium which is both untempting for player 1 and yields him a payoff worse than he would get if (A_2,B_2) were played every period involves (A_3,B_4) being played for many periods.

But is player 2 willing to play B_4 every period if she believes player 1 is playing A_3? Clearly she would benefit in the short term if she played B_2 instead. To sustain this punishment as a continuation equilibrium involves punishing a deviation by player 2. But in a RPE, if player 2 deviates after player 1 has deviated, the only continuation equilibrium permitted is the stage-game Nash equilibrium, (A_2,B_2), yielding player 2 a payoff of 10.

Thus, in any RPE, cooperation is not sustainable; some outcomes that can be supported in a SPE cannot be supported in a RPE. Note, however, that if δ were significantly closer to 1, that there would exist an RPE in which (A_1, B_1) is played every period, substainable merely by the "punishment" of (A_2, B_2). The question naturally arises of what outcomes are feasible when δ is close to 1. Does, for instance, the standard folk theorem hold?

It does not: some "individually rational" outcomes may be ruled out even as δ becomes arbitrarily close to 1. Consider Example 2 again, and consider an equilibrium yielding payoffs of less than 10 for either player. It can be verified that any equilibrium yielding less than 10 to either player, at least one of the players that is getting less than 10 can cheat. Consider such cheating by (say) player 1. Any continuation equilibrium besides (A_2, B_2) in which player 1 has no incentive to cheat yields player 2 a payoff of less than 10. But player 2 knows that if he cheats--as he must be able to do for such equilibria--he will get a payoff of 10 from now on, because they will play (A_2, B_2) from now on. Thus, player 1 knows that if he originally deviates from an equilibrium, player 2 will also deviate, yielding both players a payoff of 10. As δ approaches 1, this places a lower bound on the feasible payoffs in equilibrium.

The following theorem shows, however, that as players become patient, a large class of relatively efficient SPE outcomes can be supported in an RPE. In particular, any equilibrium in which each player gets at least his least-favorite stage-Nash payoff is consistent with RPE.

Theorem 1:

Let π_k be the payoff to player k in his lowest-payoff Nash equilibrium of the stage game G. Consider $(u_1, u_2, \ldots, u_N)$ in the convex hull of the payoffs in G such that for all k, $u_k > \pi_k$. Then for all $\varepsilon > 0$ there exists a δ^* such that, for all $\delta > \delta^*$, there exists an RPE σ in which $|v_k(\sigma) - u_k| < \varepsilon$ for all k.

Proof:

The proof involves standard folk-theorem arguments, using as "punishments" if player k cheats the continuation equilibrium where all players play k's worst stage-Nash equilibrium forever. Q.E.D.

III. Reneging, Renegotiation, and Efficiency

I now add the presumption of efficient negotiations to RPE, to consider some of the issues raised in the renegotiation literature. A naive solution concept combining RPE with efficiency is:

Definition 5:

The RPE σ is an Efficient Reneging-Proof Equilibrium (ERPE) iff there does not exist another RPE, γ, such that $v(\gamma) > v(\sigma)$.

This equilibrium concept is subject to the basic critique of the renegotiation literature: it incorporates a strong presumption of efficiency at the beginning of the game, but not at later stages. In Example 1, for instance, for a discount rate of 7/24, one of the ERPE is the cooperative equilibrium yielding the players (4,4), with the threat to play (1,1).

Recalling the discussion in the introduction, however, this seems an unlikely equilibrium in precisely those environments where negotiations are likely to be efficient: the only way to enforce cooperation is to punish deviations with a continuation equilibrium involving the repeated play of (D,R).

I now construct in several steps an alternative solution concept that incorporates Pareto-efficiency in a time-consistent manner. I begin with the crucial fact that once *all* players have lost credibility, the set of efficient continuation equilibria is simply the set of efficient equilibria involving only stage-Nash behavior. Because this set requires no punishments, it requires no theory of what are credible punishments; we need not presuppose a theory of credibility before constructing one. From this easily defined set, I use backward induction by imagining what would happen if all but one player lost credibility, if all but two players lost credibility, etc., until reaching the beginning of the game, in which no players have yet lost their credibility. (In Farrell and Maskin's [1989] terms, my approach has both "internal" consistency and "external" consistency.)

Some preliminary definitions are needed. I first define, for every subset of players Q, the set of equilibria for which no players in Q have any temptation to cheat:

Definition 6:

An RPE σ is a Q-RPE iff for all $k \in Q$, $\tau_k(\sigma) = 0$.

I next define those histories of play for which all players in Q, and only players in Q, have lost their credibility by cheating on a given RPE σ.

Definition 7:

The history b_T is a Q-history w.r.t. σ when:

A) For every player $k \in Q$, $C_k(b_T, \sigma) = 0$, and

B) For every player $k \notin Q$, $C_k(b_T, \sigma) = 1$.

Note that in any RPE, the continuation equilibrium after a Q-history must be a Q-RPE.

Consider the set of continuation equilibria that are consistent with RPE after every player has cheated, and thus nobody is trusted. If everybody has cheated, the set of equilibria consistent with the assumption of efficient renegotiation will simply be those equilibria that are Pareto-efficient within this set. Formally, let $\mathcal{N}$ be the set of all players in the game. Then $\mathcal{N}$-RPE is the set of all RPE in which no player would ever be tempted to cheat. Denote by $\mathcal{N}^{**}$-RPE those equilibria which are Pareto-efficient among the set $\mathcal{N}$-RPE.

Definition 8:

An equilibrium $\sigma \in \mathcal{N}$-RPE is in $\mathcal{N}^{**}$-RPE iff there does not exist $\gamma \in \mathcal{N}$-RPE such that $v(\gamma) > v(\sigma)$.

Now consider the case where all but player k (say) have cheated. Let $Q = \mathcal{N} \backslash \{k\}$. Then σ is in Q^*-RPE if and only if $\sigma \in Q$-RPE and, for all continuation strategies γ of σ, the payoffs from γ are not Pareto-dominated by any $\beta \in \mathcal{N}^{**}$-RPE. This restriction guarantees that if player k cheats, the continuation equilibrium will be Pareto-efficient among those equilibria are temptation-free.

Among this set of equilibria, let Q^{**}-RPE be the set of equilibria that are Pareto-efficient in Q^*-RPE. The set Q^{**}-RPE is <u>not</u> necessarily the set of equilibria that are Pareto-efficient in Q-RPE; the difference is that it rules out the use of inefficient punishments if player k deviates. This two-step process of refining the set of equilibria thus avoids the "renegotiation trap." That is, it rules out those efficient equilibria which involve the threat of inefficient punishments.

Once we define Q^{**}-RPE for every subset of $\mathcal{N}$ that excludes only one player, we can repeat the process for any subset excluding only two players, etc. We can formalize this iterative process as follows. Consider some arbitrary subset of players, Q. Then a Q-RPE σ is in Q^{*}-RPE if and only if, for every $\mathcal{R}$-history w.r.t. σ such that $Q \subset \mathcal{R}$, the continuation equilibrium of σ following that history is an $\mathcal{R}^{**}$-RPE. This definition captures the idea that every post-cheating continuation equilibrium must be efficient among the set of equilibria that are "as credible" as that continuation equilibrium. Definitions 9 and 10 combine iteratively to lead us to the main definition.

Definition 9:

A Q-RPE σ is in Q^{*}-RPE iff for all $\mathcal{R}$ such that $Q \subset \mathcal{R}$, for all $\mathcal{R}$-histories b_t, $\sigma[b_t] \in \mathcal{R}^{**}$-RPE.

This says a continuation equilibrium must have the property that, following any cheating, the players must play an equilibrium that is Pareto-efficient among the set of credible continuation equilibria. Among this set, we simply select the set of Pareto-efficient equilibria:

Definition 10:

An equilibrium $\sigma \in Q^{*}$-RPE is in Q^{**}-RPE iff there does not exist $\gamma \in Q^{*}$-RPE such that $v(\gamma) > v(\sigma)$.

Applying these definitions on iteratively smaller subsets of players starting from the set of all players, we get the following solution concept:

Definition 11:

The RPE σ is a Super Reneging-Proof Equilibrium (SRPE) iff σ is in $\emptyset^{**}$-RPE.

Theorem 2:

The set of SRPE is non-empty.

Proof:

For all Q, all equilibria consisting solely of stage-Nash behavior are in Q^{*}-RPE, because there are no deviations leading to any $\mathcal{R}$-histories for any $\mathcal{R}$'s that are strict supersets of Q. But given that Q^{*}-RPE is non-empty, Q^{**}-RPE is non-empty. Because this is true for all Q, it is true for $Q = \emptyset$, which proves the theorem.

Q.E.D.

Let us go back and consider the implications of this definition in Example 1. For $\delta \geq 1/3$, SRPE includes the symmetric equilibrium in which players cooperate by playing (U,L) until somebody cheats; after any cheating, they play (M,C) forever. For $\delta > 1/3$, there also exist SRPE involving mixed strategies over U and M and L and C, which rely heavily on the observability of strategies rather than mere actions. For $\delta < 1/3$, no SRPE involves playing (U,L) all of the time, while there may still be mixed strategy equilibria over U and M and L and C.

The solution concept SRPE can be applied to the classical games of Cournot and Bertrand competition (the infinity of equilibria in these games do not affect the basic intuition). For comparison purposes, I use the simple example of a duopoly from Farrell and Maskin [1989]. Suppose that there are two firms, Firm 1 and Firm 2, each having zero costs of production, and facing demand curve $P = 2 - q_1 - q_2$. The firms choose their strategies every year, after observing the strategies chosen in all previous years. I will discuss, a little loosely, the limit results as δ approaches 1.

Consider Cournot competition: the firms choose their quantities simultaneously, and then sell their products at the market-clearing price. What does SRPE imply? To see, let us first suppose that Firm 2 has cheated, and thus has no credibility, so that the only behavior in each period consistent with RPE are those outcomes on Firm 2's "reaction function"--i.e., those output levels for which Firm 2 has no incentive to cheat. Firm 1 knows that if *it* cheats now, it will get its Cournot outcome in all future periods. For δ close to 1, therefore, Firm 1 must get at least its Cournot payoff; otherwise it would defect. If we are interested in the set of SRPE, then we want to look at all Pareto-efficient points on Firm 2's reaction function which yield Firm 1 a Cournot payoff or better. The set of all such points involve any production by Firm 1 from its Cournot level up to its Stackelberg level (beyond which, both firms profits decline).

Considering the situation where neither firm has cheated, it is clear that Firm 2 must get at least its Stackelberg-follower payoff; otherwise it will cheat. Recognizing that Firm 1 is in the same situation, and that Pareto-efficiency involves full collusion, the set of outcomes consistent with SRPE are all those with $q_1 + q_2 = 1$, yielding total profits of 1, such that each firm gets a payoff of $\pi_i \geq 1/4$, the Stackelberg-follower payoff.

What would be the SRPE payoffs in the repeated Bertrand game with the same demand and cost functions? In the Bertrand game, if one player is not trustworthy, then both players will get payoffs of zero, because this is the only outcome at which it is not true that both firms would each benefit in the short term from deviating. Thus, in the Bertrand game, because any cheating is a disaster for both firms, the set of SRPE payoffs is equal to the set of full-collusion outcomes; all efficient outcomes are possible, and no inefficient outcomes are possible.

IV. Other Notions of Renegotiation-Proofness

We can compare RPE to the concept of Weak Renegotiation-Proofness (WRP) developed by Farrell and Maskin [1989] (thus indirectly comparing it to the related solution concepts developed by Bernheim and Ray [1989]).

I have already compared the results in Example 1, because I showed that while SRPE guarantees cooperation for discount rates $\delta > 1/3$, WRP does not allow full cooperation for any δ close to 1/3.

Consider results as δ approaches 1. In the Cournot game, WRP permits any cooperative amount yielding each firm at least a payoff of 1/8, and permits many very inefficient equilibria (including one's yielding each player their one-shot Cournot outcome). In the Bertrand game, WRP rules out some efficient equilibria (in fact, it rules out more of the efficient equilibria than it does in the Cournot game), and again permits many inefficient outcomes. SRPE guarantees efficiency, and allows all efficient outcomes.

My approach is closer to that of Pearce [1987], which I interpret as follows (but which is not necessarily compatible with the strict symmetry assumption employed by Pearce [1987]). Suppose that somebody cheats on an agreement, and some group of people, according to an agreed-upon equilibrium, are supposed to punish the deviator, but with behavior that will be costly to themselves. This gives rise to the notion of renegotiation: the group of would-be punishers might want to renegotiate with the deviator, and say that they will not punish him this time, if he agrees to behave himself. Pearce introduces history-dependent credibility by saying that a renegotiation is not credible if it involves a threat of future punishment that is just as costly as the original punishment being renegotiated away from. If the group cannot fathom a severe punishment now, can they credibly convince the deviator that they will really follow through on the punishment next time?

Pearce's approach thus focuses on the credibility of threats of future punishments, while my approach focuses on the credibility of promises to not cheat in the future. Pearce asks that society not threaten to punish criminals in the future if it has not been willing to engage in similarly costly punishment in the past; I ask that criminals not expect society to believe promises to resist the temptation of crime in the future if they have not resisted the temptation in the past.

While Pearce's assumption of symmetry may be restrictive, it does highlight an attractive feature of his model that is shared by my model, but not by the BRFM approach. The BRFM approach relies on a feature of negotiations that I shall call Pareto-indeterminacy: in a given situation, players have the option to negotiate to more than one efficient equilibrium. In order for the BRFM approach not to degenerate into an argument that behavior is independent of earlier play of the game, it must be that many efficient equilibria are possible at each stage.[3] Yet all existing models of communication which

[3] Dekel and Farrell [1990] illustrates this argument. In that paper, they argue that one-sided communication with one patient player does not lead to the Stackelberg outcome, and in fact leads only to the patient player's best stage-Nash payoff. Yet the result follows not from the fact that there is only one patient player; rather, it follows from the assumption that only one player has bargaining power, which leads to the argument that there is a

guarantee efficient equilibria also guarantee a unique equilibrium. No convincing theory of communication has yet been formulated which is both Pareto-efficient *and* Pareto-indeterminate. (And in Rabin [1991a], I argue that it might be that no such reasonable theory exists.)

It is clear that both the Pearce approach and my approach do not depend on the multiplicity of efficient continuation equilibria.

V. Two More Examples

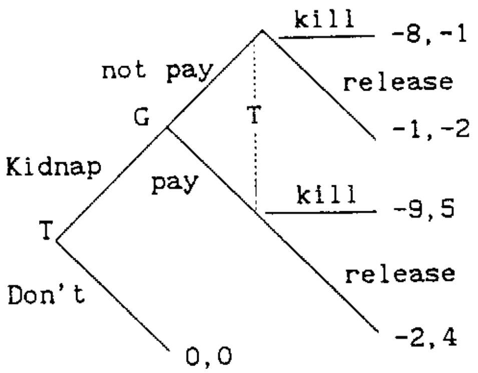
One interpretation of the renegotiation literature is that it is meant merely to "deconstruct" the efficiency hypothesis: It shows that the presumption of efficiency can lead to a contradiction in dynamic games, and imply its own opposite. Yet this contradiction is not achieved by any existing renegotiation concepts in infinitely repeated games with a discount factor close to 1. I show in this section that SRPE can rule out efficiency even as δ approaches 1. I do not wish to push the results in these examples too hard: I am not convinced the predicted behavior is uniquely plausible. But I do believe that it at least presents the possibility that the potential of efficient renegotiations may lead to an inefficient equilibrium.

Before providing an example where inefficiency is guaranteed, I first illustrate how the issue of credibility can rule out some efficient equilibria. Efficient equilibria can be ruled out by an interesting incentive created in some games: a player might wish to "signal" his untrustworthiness,

unique Pareto-efficient continuation equilibrium at each point. In a repeated game with one-sided communication, the same non-cooperation argument would obtain. Indeed, any theory of renegotiations which is Pareto-determinate would degenerate into non-cooperation.

because his payoff may be higher if he is not trusted than if he is. Consider Example 3.[4]

Example 3



		Terrorists		
		Don't	Kill	Release
Govt.	Don'tpay	0,0	-8,-1	-1,-2
	Pay	0,0	-9,5	-2,4

In this example, a terrorist organization considers each period whether or not to kidnap some citizens of a nation, and then demand a monetary or political payoff in exchange for the release of the hostages. Then, simultaneously (or without observing each other's decision), the government chooses whether to pay the terrorists, and the terrorists choose whether to release the hostages. The government wants no terrorist activity, prefers the release of hostages to their death, and prefers to not pay rather than to pay. The terrorists prefer to kidnap people if and only if they will get paid by the government, and prefer to kill the hostages to releasing them (this last assumption is not central).

There are many equilibria to the infinitely repeated game, including equilibria in which the terrorists kidnap citizens each period, and release them in exchange for payment by the government. This is supported by threats that if the government does not pay, the terrorists will in the future kill

[4] This example was suggested by Joe Farrell.

some hostages.

The only outcome consistent with SRPE as the discount factor approaches 1, however, is that the terrorists never kidnap. This is because, if the terrorists ever kidnap under the understanding that the government will pay, the government can "cheat" by refusing to pay. Once the government does this, it signals that it cannot be trusted by the terrorists in the future. If the terrorists do not trust the government, then they will never kidnap.

In Example 3, the desire by one player to signal its propensity to cheat led to a unique Pareto-efficient outcome in the repeated game. Example 4 shows that, when *both* players wish to signal their propensity to cheat, SRPE can rule out *all* Pareto-efficient subgame-perfect equilibria:

Example 4

			Player 2	
		L	C	R
Player 1	U	1,1	9,0	4,0
	D	0,9.9	10,2	2,10

In this game, there is a unique stage-Nash equilibrium, (U,L), with payoffs (1,1). Although there are subgame-perfect equilibria with payoffs that Pareto-dominate (1,1), it turns out that (1,1) is the unique SRPE outcome.

To see this, consider the set of Pareto-efficient payoffs in the case where player 1 has lost his credibility, and player 2 has not. The only play in the stage game for which it is not tempting for player 1 to deviate is either (U,L) or some mixed strategy in which player 1 is indifferent between playing U and D. The lowest payoffs among such mixed strategies yields player

1 a payoff of 5, and player 2 a payoff above 1. Thus, if players play an equilibrium that is Pareto-efficient among those in which player 1 is not trusted, it will yield player 1 a payoff of 5 or more.

Suppose player 2 has lost credibility. Then the only strategy pairs in which player 2 has no temptation to cheat are either (U,L), or strategies where player 1 mixes between U and D and player 2 mixes between L and R. Among this set of strategy pairs, any Pareto-efficient equilibrium must yield at least 2 to player 1 and at least 9 to player 2.

Thus, if player 2 is the first to cheat, he gets a minimum payoff of 9; if player 1 is the first to cheat, player 1 gets a minimum payoff of 5. Thus, in order for an equilibrium to be an SRPE, it must be that one of the following is true: (A) player 1 gets at least 5, player 2 gets at least 9; (B) player 1 gets at least 5, and player 2 *cannot* cheat; (C) player 2 gets at least 9, and player 1 *cannot* cheat; or (D) neither player can cheat. Case (A) cannot occur because no outcomes in this game yield such payoffs. Case (B) cannot occur because any outcome that yields player 1 a payoff of 5 or more is also tempting for player 2. Case (C) cannot occur because any outcome which yields a payoff of 9 or more for player 2 is tempting for player 1. Only (D) is possible, and can occur only if the players play (U,L) every period. Thus no cooperation can occur, because any cooperation tempts one player or the other to cheat so as to lose credibility, and get himself a better outcome. Efficiency is ruled out.

Is this realistic? While I do not want to oversell the outcome, it seems at least plausible. If both players benefit from not being trusted, they will distrust any outcome where the other player is tempted to cheat.

What is problematic, however, is that player 1 (say) would do better by reaching an agreement that he knows player 2 will cheat on than he does in the unique SRPE outcome. He is quite happy to lose trust in player 2. Yet it is not obvious that this should rule out the SRPE outcome. Credibility should probably not be interpreted solely in terms of a player "cheating" on an explicitly negotiated agreement, if it is *common knowledge* that he will cheat. If cheating is commonly expected, the players may treat the expected behavior--rather than the proposed equilibrium--as the "true" agreement. From this perspective, the only outcome that can be commonly expected in any period of play in Example 4 is (1,1).

V. Conclusion

I have assumed throughout that cheating once, to whatever degree, renders a player untrustworthy for life; giving in to temptation once yields eternal damnation. This is certainly an extreme assumption, and in polar contrast to the BRFM assumption that no loss of credibility accrues from cheating.[5]

While I suspect that many of the results from SRPE would hold with milder forms of loss of credibility, the extreme assumptions I make may be more worrisome in more general games. As with Bernheim and Ray [1989] and Farrell and Maskin [1989], I define my solution concepts for games with perfect observability after each stage game. In contrast, Pearce [1989] and Abreu, Pearce, and Stacchetti [1989] consider the case where there is imperfect

[5] One candidate for a less dramatic loss of credibility from cheating is to suppose that a player loses credibility just long enough so that the cheating was not worth his while. In this case, if a player has genuinely "slipped", rather than cheated, his credibility will eventually be restored. While this way of limiting loss of credibility is certainly contrived, it indicates that there is a crucial cutoff point for how much distrust is generated by cheating; if players are distrustful enough of cheating players, the results of this paper will obtain.

observability between periods. If there is imperfect observability, how would players judge whether another player has cheated or not? Either they would wait for accumulated evidence and punish severely, or they would have to punish mildly but frequently if suspicious outcomes occur that cannot be attributed for sure to cheating. A reformulation of my solution concept would have to contend with such issues, where "punishments" occur even when no true cheating has occured, and when it is common knowledge among the players that no cheating has occured.[6]

I close with a critique of the entire renegotiation-proofness literature, of which this paper is a part. All the stories used to justify solution concepts rely on arguments about plausible interpretations of "deviations." As many game theorists are beginning to realize, however, a "complete" theory ought be explicit about what causes deviations. Perhaps the most natural interpretation in repeated games is to suppose that the games are never really of complete information. Theories of renegotiation should then be based on incomplete-information models, where different theories of renegotiation may correspond to different theories of what types of "deviant" players tend to "infect" most games.[7]

Yet whatever the outcome of formal game-theoretic research, it is hard to imagine many real-world situations in which repeated deviations by one person from agreed-upon behavior does not eventually lead to distrust in that person.

[6] Note that this problem arises in some games even when players can observe perfectly all previous actions, but cannot directly observe (mixed) strategies.

[7] For instance, it might be that SRPE corresponds approximately to something like the theory that all games are infected with "myopic" players, who have the same per-period payoffs as the explicitly modeled players, but have a discount factor of zero. Such types would always "cheat" (and would never slip). I have not attempted to model this formally.

References

Abreu, Dilip and David Pearce (1989), "A Perspective on Renegotiation in Repeated Games," Harvard Institute of Economic Research Discussion Paper 1453.

Abreu, Dilip, David Pearce and Ennio Stacchetti (1989), "Renegotiation and Symmetry in Repeated Games," mimeo, Yale University, May.

Benoit, Jean-Pierre, and Vijay Krishna (1988), "Renegotiation in Finitely Repeated Games," Harvard Business School Working Paper 89-004

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

Dekel, Eddie and Joseph Farrell (1990), "One-Sided Patience with One-Sided Communication Does Not Justify Stackelberg Equilibrium," *Games Econ. Behav.* **2**, 299-303.

Evans, R. and Eric Maskin (1989), "Efficient Renegotiation-Proof Equilibria in Repeated Games," *Games Econ. Behav.* **1**, 361-369.

Farrell, Joseph and Eric Maskin (1989), "Renegotiation in Repeated Games," *Games Econ Behav.* **1**, 327-360.

Pearce, David (1987), "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," mimeo, Yale University.

Rabin, Matthew (1991), "A Model of Pre-game Communication," mimeo, Berkeley, April.