GAMES AND ECONOMIC BEHAVIOR 2, 299–303 (1990)

One-Sided Patience with One-Sided Communication Does Not Justify Stackelberg Equilibrium*

EDDIE DEKEL AND JOSEPH FARRELL

Department of Economics, University of California, Berkeley, California 94720

Received February 13, 1990

The theory of games (with complete information) in which a single patient long-run player faces a succession of short-run opponents cannot plausibly be used to justify the Stackelberg solution concept, because if that player can select which subgame-perfect equilibrium is to be played then she can presumably also change her selection. Consequently, while she can choose among one-shot Nash outcomes, she cannot achieve the Stackelberg outcome. *Journal of Economic Literature* Classification Number: 026. © 1990 Academic Press, Inc.

The Stackelberg equilibrium concept is completely natural in a one-shot game in which one player, the "leader" L, moves first, and the other observes L's move before choosing his own; indeed, in this case it is no more than subgame-perfect equilibrium. But in many applications there is no such sequential structure: instead, a simultaneous-move game is infinitely repeated and the solution "play repeatedly the Stackelberg equilibrium in the stage-game" is felt to be justified because one (and only one) player, L, can somehow "commit" over time to a particular stage-game action. In particular, it may seem intuitively appealing that if player L is long-lived and has a discount factor close to 1, while her opponent has a very low discount factor (or is represented by a sequence of short-run players), then L will be able to take the role of Stackelberg leader, and in each period the one-shot Stackelberg outcome, s, will occur.

There are several possible ways in which one might hope to formalize this intuition. First, Fudenberg and Levine (1989) show that if there is incomplete information of a certain kind about the long-run player's pay-

* We thank an associate editor for his comments, and the National Science Foundation (Grants SES 88 08133 and IRI 87 12238) for financial support.

0899-8256/90 $3.00

offs, then she will achieve an overall payoff close to that from s being played each period. More precisely, say that the long-run player is a "Stackelberg type" if it is a strictly dominant strategy for her to play her Stackelberg action s_L in each period. Fudenberg and Levine show that if the short-run players assign strictly positive probability to the long-run player being a Stackelberg type, then in any Nash equilibrium the long-run player will achieve a payoff close to that from s being played each period. The result follows from a "reputation effects" argument (Kreps *et al.*, 1982; Fudenberg and Maskin, 1986): the long-run player will mimic the Stackelberg type. But, for this justification, we must assume that the short-run players believe that with positive probability the long-run player is a Stackelberg type. Although it is often plausible that there is incomplete information about the precise payoffs, we think it is worth considering situations in which it *is* common knowledge that the long-run player is *not* the Stackelberg type.

In this note, we examine another method, which rests on the claim that even with complete information, the long-run player can credibly commit to any form of behavior, and will therefore choose to commit to the most profitable form of behavior, i.e., playing s_L always. More precisely: (i) there is a subgame-perfect equilibrium in which the long-run player always plays s_L; moreover, (ii) this yields her best payoff among subgame-perfect equilibria. Intuitively, one might therefore think that (iii) if she controls the communication she will select this equilibrium. Parts (i) and (ii) of this claim have been justified by Fudenberg *et al.* (1988), as we discuss below. Part (iii), however, is invalid, as we show.

Fudenberg *et al.* (1988) characterized the subgame-perfect equilibria of games in which one player, who has a discount factor δ close to 1, plays against a succession of "short-run" players, each of whom plays the game only once (although he observes the entire history).[1] They derive a "Folk Theorem" result as follows. Restrict attention to action-pairs in which the short-run player plays a best response to the long-run player's move. Defining feasibility and minimax payoffs relative to this restricted set of action-pairs, they show that, as $\delta \to 1$, all feasible, strictly individually rational, outcomes become subgame-perfect equilibria. In particular, if the long-run player's mixed strategies are observable, then there exists a subgame-perfect equilibrium in which along the equilibrium path s is played in each period.[2] Abusing terminology, we call this subgame-perfect equilibrium "the Stackelberg equilibrium." Fudenberg *et al.* also show

[1] In fact, they allowed for more than one long-run player and for more than one short-run player, but we restrict our attention to the simplest case.

[2] The reader is referred to their paper for a discussion of the subtle issues that arise if mixed strategies are not observable.

that there is no subgame-perfect equilibrium that the long-run player strictly prefers to this equilibrium.

To conclude from this that the Stackelberg equilibrium may be expected in such a game, we must also assume that the long-run player can choose which subgame-perfect equilibrium will be played. Intuitively, this may seem reasonable: after all, she has in some sense the most at stake, and might well be expected to control the channels of communication. If so, then it seems plausible that she can select her preferred equilibrium.[3] In any case, whether plausible or not, the assumption is necessary in order to justify Stackelberg equilibrium in this fashion.

But if the long-run player can *select* an equilibrium at the beginning of the game, it is natural to suppose that she can also *reselect* later. This observation suggests an analogy to the theory of renegotiation-proof equilibrium.[4] There, it is supposed that players *jointly negotiate* an equilibrium of the repeated game, but cannot commit themselves not to *renegotiate* later, if doing so would yield a strictly better continuation outcome for all of them. Here, we suppose that one player can *unilaterally select* an equilibrium at the beginning; unless the opportunities for communication occur only once, we must also suppose that she can *reselect* later, if doing so would yield a strictly better continuation outcome for her.

Farrell and Maskin (1989) call a subgame-perfect equilibrium "weakly renegotiation-proof" if no two of its continuation equilibria are strictly Pareto-ranked. In the same spirit, we call a subgame-perfect equilibrium "weakly reselection-proof" if no two of its continuation equilibria are strictly ranked according to the long-run player's payoffs. Like weak renegotiation-proofness, this seems *necessary* for credibility: indeed, the assumption that a player who controls communication can achieve her best equilibrium seems even more compelling than the assumption that players jointly can always agree on, and thereby achieve, Pareto improvements.

But, unless it happens also to be a one-shot Nash equilibrium, the Stackelberg equilibrium is *not* weakly reselection-proof. To see this, note that in any weakly reselection-proof equilibrium, every two continuation equilibria must yield the same continuation payoff to the long-run player. Therefore she cannot be threatened by future consequences of her current opportunism, and so she will always choose short-run best responses. Consequently, any weakly reselection-proof equilibrium must specify, in every period, a one-shot Nash equilibrium of the stage-game.

[3] See Farrell (1988) for a formal argument in the case of one-shot games.

[4] Bernheim and Ray (1989) and Farrell and Maskin (1989) independently developed the theory reviewed below. See Pearce (1988) for a different approach to renegotiation, and Benoit and Krishna (1988) for the finitely repeated case.

We conclude that the Fudenberg–Kreps–Maskin theory of games with one patient (long-run) player facing one or more short-run players cannot justify the Stackelberg equilibrium concept, absent some explanation of why the long-run player has selection power but not reselection power.

Because weak reselection-proofness is a necessary condition, we can rule out subgame-perfect equilibria, such as the Stackelberg equilibrium, that fail to satisfy it. But it is by no means a sufficient condition: for example, suppose that the stage-game has two Nash equilibria, e_1 and e_2, and that the long-run player strictly prefers e_2 to e_1. Then "always play e_1, regardless of history" is a weakly reselection-proof equilibrium (it has no continuation equilibria other than itself), but it is plainly implausible.

This leads us to define a "strongly reselection-proof" equilibrium (by analogy with strongly renegotiation-proof equilibrium) as a weakly reselection-proof equilibrium no continuation equilibrium of which is strictly dominated, for the long-run player, by *any* weakly reselection-proof equilibrium. Like strongly renegotiation-proof equilibrium, strongly reselection-proof equilibrium seems convincing if it exists.

But it is evident that strongly reselection-proof equilibrium always exists, and that such equilibria are precisely the weakly reselection-proof equilibria in which each period's one-shot Nash equilibrium gives the long-run player her maximum one-shot Nash equilibrium payoff.[5] This formalizes the idea that the long-run player can choose among one-shot Nash equilibria, although (because she cannot commit not to reselect) she can do no better than that.

REFERENCES

BENOIT, J.-P. AND KRISHNA, V. (1988). "Renegotiation in Finitely Repeated Games," Harvard Business School Working Paper 89-004.

BERNHEIM, B. D., AND RAY, D. (1989). "Collective Dynamic Consistency in Repeated Games," *Games Econ. Behav.* **1,** 295–326.

FARRELL, J. (1988). "Communication, Coordination, and Nash Equilibrium," *Econ. Lett.* **27,** 209–214; see also forthcoming *Erratum.*

FARRELL, J., AND MASKIN, E. (1989). "Renegotiation in Repeated Games," *Games Econ. Behav.* **1,** 327–360.

FUDENBERG, D., KREPS, D. AND MASKIN, E. (1988). "Repeated Games with Long-Run and Short-Run Players," MIT Working Paper 474.

FUDENBERG, D., AND LEVINE, D. (1989). "Reputation and Equilibrium Selection in Games with a Patient Player," *Econometrica* **59,** 759–778.

[5] For a typical stage-game, of course, there is just one such one-shot Nash equilibrium, so there is a unique strongly reselection-proof equilibrium.

FUDENBERG, D., AND MASKIN, E. (1986). "The Folk Theorem in Repeated Games with Discounting, and with Incomplete Information," *Econometrica* **54,** 533–554.

KREPS, D., MILGROM, P., ROBERTS, J., AND WILSON, R. (1982). "Rational Cooperation in the Repeated Prisoner's Dilemma," *J. Econ. Theory* **27,** 245–252.

PEARCE, D. (1988). "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," mimeo, Yale.