Econometrica, Vol. 60, No. 6 (November, 1992), 1435–1438

NOTES AND COMMENTS

A NOTE ON ABREU-MATSUSHIMA MECHANISMS

BY JACOB GLAZER AND ROBERT W. ROSENTHAL[1]

IN A STIMULATING RECENT PAPER, Abreu and Matsushima (1992) (hereafter A-M) show how a class of social choice functions can be virtually implemented in iteratively undominated strategies. This work has several important features: the mechanisms used are finite and not too difficult to understand, and so less objectionable in this regard than many in the literature; the class of social choice functions implemented is large; and the solution concept—Nash equilibrium determined uniquely by iterative elimination of strongly dominated strategies—is relatively uncontroversial. The point of this note is to argue that the mechanisms used by A-M unfortunately tend to generate games in which the iterative removal of strongly dominated strategies sometimes is indeed (or ought to be) controversial.[2]

We proceed by first examining an example of a related but simpler implementation problem in which the argument is easily exposed, then indicating how the argument applies generally in the A-M setup.

Consider the following much-discussed two-player coordination game:

	α	β
α	1,1	0,0
β	0,0	2,2

This game has the obvious two pure-strategy Nash equilibria, plus a mixed-strategy equilibrium that yields the expected payoffs $(2/3, 2/3)$. Most commentators argue that the (β, β) strategy combination is the obvious choice on focal-point grounds, but that the others cannot be ruled out by using standard refinements.

Imagine now that some third party (Ms. P) has an interest in getting the players of this game to play (α, α). To accomplish her task, P has two tools at her disposal. First, she can levy fines (denominated in utils scaled to conform with the game's payoffs); but the amounts of these must be kept small relative to the magnitudes of the payoff differences in the game. Second, she can force the players to play the game in very small pieces. Specifically, suppose P sets out the following common-knowledge rules:

The game is played in 100 stages. For each stage each player chooses from the possible randomizations over $\{\alpha, \beta\}$. All 100 of both players' choices must be made simultaneously at the beginning of the game, however, so selections cannot be functions of the history of play. The overall expected payoff for each player is simply the average of his expected payoffs in the stages. From each player's overall expected payoff, a fine of .05 utility units is subtracted if the first stage in which that player's choice is β occurs no later than the stage of the first β choice of his co-player. There are no other fines.

With this mechanism, P accomplishes her goal in the sense that the only strategy combination that survives iterative removal of strongly dominated strategies (and hence the unique Nash equilibrium) is for both players to produce lists of 100 α's. To see this, note first that any strategy that calls for α in the first stage strongly dominates that same

[1] We are grateful to Andrew Weiss for stimulating discussions on this subject. The second author thanks the National Science Foundation for financial support under Grant #SES-9010246.

[2] Similar mechanisms are used in Abreu and Matsushima (1990a, 1990b). Similar criticisms can be applied. We spare the readers the details.

strategy with anything else (β or any nondegenerate mixture) in the first stage: If the co-player's first-stage realization is β, α loses .02 relative to β in the overall game but avoids the .05 fine altogether. If the co-player's first-stage realization is α, α gains .01 relative to β in the overall game and avoids the .05 fine temporarily. In this case, even if the fine is ultimately paid there is still left the gain of .01. Hence we may remove all strategies that involve a first-stage choice of anything other than α with probability one and proceed by induction to the rest of the stages.

By way of interpretation, P's diabolical selection of rules forces players who accede to the logic of iterated dominance to focus not only on the trees (the stages) instead of the forest (the overall game) but on the trees in sequence, so that the small fines at her disposal can be threatened again and again, but never actually used, to kill off the focal equilibrium in the overall game. The trouble with this is that the focal-point story that leads to the selection of $\{\beta, \beta\}$ in the first place plays no role in the iterated-dominance logic and so can have no countervailing influence there on the players' choices. We suspect, however, that contemplation of the forest—that one or two units of utility are at stake in total—would lead players to abandon the logic of iterated dominance in favor of the focal point in this game much as it would (and seems to in experiments) in the finitely-repeated prisoner's dilemma, where the iterated-dominance logic would force the players to double-cross at every stage.[3]

Before proceeding to the A-M mechanism itself, we note that with the same tools—small fines (that are never actually paid) and pieced play—and in the same sense unique strategy combination that survives iterative removal of strongly dominated strategies—P can implement any Nash equilibrium of any finite noncooperative game in almost the same way. To the mechanism analogous to the one above must be appended an arbitrarily small fine in each stage for each player whenever that player departs from his part of the equilibrium strategy combination. To see why this is needed, change the payoff to strategy combination (α, α) in the example from $(1,1)$ to $(0,0)$. Now the first-stage strong dominance argument fails for the case when the co-player plays α. The extra fine is used to break such indifference whenever the Nash equilibrium being implemented is not a strict equilibrium.

The A-M mechanism applies to a broad class of problems of a different sort than the example above, but logic analogous to the above is central in that mechanism's design, and the criticism above applies to an important subset of the problems in the domain under consideration in that paper. A-M study the following virtual implementation problem. There are $N \geqslant 3$ agents, each of whom is endowed with a von Neumann-Morgenstern utility function over a finite set of social states. All agents are assumed to know the entire profile of utilities. A social planner (Ms. P again) who is ignorant of the preferences of the agents wishes to implement a social-choice function that maps the set of possible utility profiles into the set of lotteries over the social states. A-M show that for every $\varepsilon > 0$, P can design a mechanism such that the unique outcome of the game induced by this mechanism (where induced outcome is defined as a strategy combination that survives iterative elimination of strictly dominated strategies) is one in which P chooses with probability at least $(1-\varepsilon)$ the state assigned to the true utility profile by the social-choice function.

In order to see how the A-M mechanism works, consider the following simple example: There are 3 agents, all of whom are either of type a or of type b (i.e., the types are perfectly correlated across agents). The agents know their common type; P does not. P must choose between the two social states α and β. Let $u(i,j)$ denote the utility of

[3] There are differences between the finitely-repeated prisoner's dilemma and the game discussed here. We mention that example just to remind the reader that the iterated-dominance logic has been called into question before. See, for example, pp. 393–399 in Kreps (1990).

each agent of type i when P chooses the social state j. Assume that

$$u(a,\alpha)=u(b,\beta)=2$$

and

$$u(a,\beta)=u(b,\alpha)=1.$$

That is, if the agents are of type a they prefer that P chooses the alternative α and if they are of type b they prefer that P chooses the alternative β. Suppose, however, that P's objective is to do the opposite. That is, when the players are of type a she wants to choose the alternative β and when the players are of type b she wants to choose α. Suppose in addition that P can fine each of the agents in total no more than some small amount $\delta > 0$. Let a small positive ε be given; set $\gamma = \min\{\varepsilon/6, \delta/2\}$, and let K be an integer such that $K > 1/\gamma$. P sets up the following mechanism, the outcome of which is that with probability $(1-\varepsilon)$ P chooses the social state preferred by her and with probability ε she chooses the social state preferred by the agents:

Each agent submits a list of $K+1$ announcements indexed by $k = 0, 1, \ldots, K$. The lists of all agents are submitted simultaneously. In each of his announcements the agent reports his (and, hence, the others') type. P then uses these lists to choose a lottery that randomizes between the two alternatives in the following way. From a starting point having probability zero assigned to both states, she first looks at the zeroth announcement of each agent and adds probability $\varepsilon/3$ to alternative α for each agent who reported a and probability $\varepsilon/3$ to alternative β for each agent who reported b. Then, for each announcement $k \geqslant 1$, P adds probability $(1-\varepsilon)/K$ to alternative β if at the kth announcement at least two agents reported a and adds probability $(1-\varepsilon)/K$ to alternative α otherwise. P fines the agents under two circumstances. If at some announcement $k \geqslant 1$ an agent reports one type while at all smaller values of k he and all the other agents had reported the other type, the agent is fined the amount γ. In addition, each agent is fined the amount γ/K for each announcement $k \geqslant 1$ in which the other two agents report one type and he disagrees.

Given this mechanism it is easy to see that any strategy (list) in which an agent does not report his true type at announcement 0 is strictly dominated by the same strategy modified so that at announcement 0 he reveals his true type. However, after elimination of all these dominated strategies for all agents it then becomes a strictly dominated strategy for an agent not to report his true type at announcement 1. Continuing in a similar way, the only strategy for every agent that survives iterative elimination of strictly dominated strategies is the one where the agent reports his true type at all $(K+1)$ announcements. The planner ends up choosing with probability ε the state preferred by the agents and with probability $(1-\varepsilon)$ the state preferred by herself.

The main differences between this example and the first one are the presence of ε, the zeroth stage, and the majority rule. The first two force an initial truthful revelation, given what is to come, in return for a small chance at the agents' preferred state. P then uses a variation of the scheme in the first example to get the agents to continue in such a way that they knowingly produce her preferred state with all the rest of the probability, although the other announcement is a countervailing focal point for them. Thus our misgivings about the mechanism in the first example apply equally well to this example.

The A-M mechanism works in general much the same as the scheme in our second example. It accomplishes its virtual (ε)-implementation goal for a broad class of social-choice functions and does not use fines,[4] but at its heart is a game like the ones above in which the iterated-dominance logic produces the desired result whether or not there is a

[4] Abreu and Matsushima themselves introduce fines as a way to explain the logic of their mechanism informally.

countervailing focal point. Such a focal point will always be present whenever the social-choice function fails to satisfy the Pareto criterion with respect to the agents' true preferences. When the social-choice function does satisfy Pareto optimality, our objection loses much of its immediate force; but we would still hesitate to give long odds on P's desired state being chosen by subjects in carefully controlled experiments.

Dept. of Economics, Boston University, Boston, MA 02215, U.S.A., and Faculty of Management, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel

and

Dept. of Economics, Boston University, Boston, MA 02215, U.S.A.

Manuscript received April, 1991; final revision received March, 1992.

REFERENCES

ABREU, D., AND H. MATSUSHIMA (1990a): "Virtual Implementation in Iteratively Undominated Strategies II: Incomplete Information," Working Paper, Princeton University.

——— (1990b): "Exact Implementation," Working Paper, Princeton University.

——— (1992): "Virtual Implementation in Iteratively Undominated Strategies I: Complete Information," *Econometrica*, 60, 993–1008.

KREPS, D. (1990): *A Course in Microeconomic Theory*. Princeton: Princeton University Press.