GAMES AND ECONOMIC BEHAVIOR **15,** 27–32 (1996)
ARTICLE NO. 0057

Virtual Implementation in Backwards Induction

Jacob Glazer

The Faculty of Management, Tel Aviv University, Tel Aviv, Israel

and

Motty Perry

Department of Economics and Center for Rationality, The Hebrew University of Jerusalem, Jerusalem, Israel

Received May 12, 1993

We examine a sequential mechanism which is a simple modification of the normal form mechanism introduced by Abreu and Matsushima (1992). We show that almost any social choice function can be virtually implemented via a finite sequential game of perfect information. The solution concept assumed is backwards induction. In particular, any social choice function that is virtually implementable via the Abreu–Matsushima mechanism is also virtually implementable by a sequential mechanism. *Journal of Economic Literature* Classification Number: C72. © 1996 Academic Press, Inc.

INTRODUCTION

Implementation theory has come a long way since the pioneering contributions of Hurwicz (1972), Gibbard (1973), Satterthwaite (1975), Maskin (1977), and others. The fundamental question addressed in this literature is that of which social choice functions are implementable and under what assumptions. While the first results in this area were mostly negative (e.g., Satterthwaite, 1975, and Gibbard, 1973, for implementation in dominant strategies), research has taken a somewhat different and more positive direction in recent years. Starting with Maskin (1977), who gave necessary and sufficient conditions for Nash implementation, researchers have studied implementation problems under various solution concepts. See Moore (1991) and Repullo (1990) for a comprehensive survey of this literature.

However, most of the mechanisms suggested by the literature were very com-

0899-8256/96 $18.00

plicated and relied on unlikely actions taken by the agents. It was recognized early on that if implementation theory was to have any relevance in real-life applications, researchers would have to focus their efforts, looking for mechanisms that are simple and, at the same time, supported by a convincing solution concept. (See, for example, Jackson (1992).)

In a recent stimulating paper, Abreu and Matsushima (1992) made an important step in this direction. They showed that almost any social choice function is virtually implementable (see also Abreu and Sen, 1991, for earlier results on this issue). Moreover, the mechanism they suggested, hereafter referred to as the A–M mechanism, was very simple and utilized a very strong solution concept, iterative elimination of strictly dominated strategies. Thus, by moving from exact to virtual implementation, a relatively minor loss, the mechanism is both simplified and strengthened significantly.

Our paper goes one step further in simplifying the mechanism and strengthening the solution concept. In Abreu and Matsushima (1992) it is essential that players do not observe each other's reports upon submitting their own reports. Thus, players must move simultaneously. We examine which social choice functions can be virtually implemented when players cannot make simultaneous moves. In other words, we analyze a finite game of perfect information in which players' reports cannot be hidden. (These mechanisms are sometimes referred to as sequential mechanisms.)

We show that almost any social choice function can be virtually implemented via a sequential mechanism, where the solution concept assumed is subgame perfect equilibrium. (*Note*. We use the terms "subgame perfect equilibrium" and "backwards induction" interchangeably.) In particular, any social choice function that is virtually implementable via the A–M mechanism is also virtually implementable by a sequential mechanism.

Sequential mechanisms are interesting for several reasons. First and foremost, sequential mechanisms, with backwards induction as their solution concept, seem to be more intuitive and simpler to understand than their simultaneous counterparts. Second, most of the mechanisms in which players are asked to submit their reports simultaneously will not work if, instead, some of the players move sequentially, each getting to observe all reports of all the agents preceding him, before submitting his own report to the planner. In many real-life situations, it is difficult for the planner to ensure "secrecy" (especially in cases where it is in the players' interest to reveal their information). On the other hand, the planner can always commit to asking the players to submit their report sequentially and to reveal each report's content upon its arrival.

Finally, on a theoretical level, one may ask whether the restriction of allowable mechanisms to sequential ones reduces the set of social choice functions that can be implemented and, if so, to what extent.

The sequential mechanism we suggest is a simple modification of the normal form mechanism introduced by Abreu and Matsushima (1992). Although we have attempted to keep our model self-contained, we would encourage the reader who has not yet done so to first read their paper before studying ours.

SETUP

Let $N = \{1, 2, \dots, n\}$ denote the set of players and denote by Ψ_i a finite set of types for player i, $i \in N$. The set of pure social alternatives is denoted by A, and $\Delta(A)$ denotes the set of all probability distributions over A. In this context, $a \in A$ denotes a pure social alternative and $l \in \Delta(A)$ denotes a lottery on A.

The utility function of player i over the set A is denoted by $u_i\colon \Psi_i \times A \to \mathbf{R}$, where $u_i(\psi_i, a)$ specifies the utility of player i from the social alternative a, when he is of type ψ_i. Player i's utility from a lottery $l \in \Delta(A)$ is $U_i(\psi_i, l) = E_l[u_i(\psi_i, a)]$. We assume that u_i is not a constant function and that for any $\psi_i', \psi_i'' \in \Psi_i$, $\psi_i' \neq \psi_i''$, $u_i(\psi_i', \cdot)$ is not a linear transformation of $u_i(\psi_i'', \cdot)$.

A social choice function is a mapping $X\colon \Psi \to \Delta(A)$, where $\Psi = \Psi_1 \times \Psi_2 \times \cdots \times \Psi_n$.

For simplicity we assume a full domain of preferences. Assume that the preferences profile $\psi \in \Psi$ is common knowledge among the players. A sequential mechanism is a complete information game form, denoted by (Γ, g), where g assigns to each strategy profile in the tree Γ a probability distribution $l \in \Delta(A)$.

THEOREM. *Suppose that $n > 2$. Then, for any social choice function X and $\varepsilon > 0$, there exists a sequential mechanism for which the unique subgame perfect equilibrium is such that for every profile of types ψ, the alternative $X(\psi)$ is chosen with probability of at least $1 - \varepsilon$.*

Proof. Assume some social choice function X and some $\varepsilon > 0$. The proposition is proved by construction of a mechanism.

To simplify the exposition we assume that the planner can fine every player i, $i \in N$ by an amount $t_i \in \mathbf{R}$, where $t_i \leq \bar{t}$, for some $\bar{t} > 0$. We also assume that player i's **VNM** utility from a transfer t_i and a lottery $l \in \Delta(A)$ is $U_i(\psi_i, l, t_i) = E_l[u_i(\psi_i, a)] - t_i$. We restrict ourselves to mechanisms in which the only money transfers allowed are from the players to the planner. The introduction of fines is made for the sake of simplicity. We shall return to this point after presenting the mechanism. It is important to notice that we make no assumption about the size of $\bar{t}$, except that it is greater than zero.

A sequential mechanism, in this setup, is a complete information game form (Γ, g), where g assigns to each strategy profile in Γ a pair (l, t), where $l \in \Delta(A)$, $t = (t_1, t_2, \dots, t_n)$, and $0 \leq t_i \leq \bar{t}$.

The Sequential Mechanism. The game is played in $K + 1$ stages. In each stage h, $h = 1, \dots, K$, each of the n players announces a vector $m_i^h \in \Psi$, which is a profile of types for the n players. In stage $K+1$ each player i announces only a type for himself, $m_i^{K+1} \in \Psi_i$. Let $m^{K+1} = (m_1^{K+1}, m_2^{K+1}, \ldots, m_n^{K+1})$, be the profile obtained from the messages in stage $K+1$. In each of the $K+1$ stages the players move *sequentially*, player 1 moves first, player 2 moves second, etc. Each player is aware of all previous announcements of all players (i.e., all announcements made in previous stages as well as those made by previous players in the current stage). Thus, this is a game of *perfect information*.

We can now discuss the planner's choice of alternative and fines given a play of the game. The planner's decision is similar to the one in the A–M mechanism with two modifications, (i) there exists only *one* type of punishment (instead of the two as in A–M), and (ii) the *last* player to disagree with the profile obtained from the $(K+1)$-st announcement is punished (instead of the first one in A–M).

For each player $i \in N$, construct a function $f_i\colon \Psi_i \to \Delta(A)$ such that for all $\psi_i \in \Psi_i$,

$$U_i(\psi_i, f_i(\psi_i)) - U_i(\psi_i, f_i(\psi_i')) > 0 \qquad \text{for all } \psi_i' \neq \psi_i.$$

Such a function always exists, in this setup, as was shown by Abreu and Matsushima (1992).

Choose some Δ so that

$$0 < \Delta < U_i(\psi_i, f_i(\psi_i)) - U_i(\psi_i, f_i(\psi_i')) \qquad \text{for all } i, \psi_i \text{ and } \psi_i' \neq \psi_i.$$

Let

$$\xi = \max_{i\in N;\psi_i\in\Psi_i;\psi',\psi\in\Psi} [U_i(\psi_i, X(\psi)) - U_i(\psi_i, X(\psi'))],$$

Choose an integer K and a fine $\delta > 0$ such that

$$\xi/K < \delta < \min\{\varepsilon\Delta/n, \bar{t}\}.$$

We are now ready to specify the function g that assigns for each strategy profile in Γ a pair $(1, t)$. For each stage h, $h = 1, \ldots, K$, a probability of $(1-\varepsilon)/K$ is assigned to $X(\psi)$ if $m_i^h = \psi$, for at least $n-1$ players; otherwise a probability of $(1-\varepsilon)/K$ is assigned to some arbitrarily chosen alternative b. In addition, for $i = 1, \ldots, n$, a probability of ε/n is assigned to $f_i(m_i^{K+1})$.

Finally, the mechanism fines at most one player: If player i is the last one to disagree with the profile obtained in stage $K+1$, then player i is fined by δ, i.e., if and only if, for some $h \in 1, \ldots, K$ and $i \in N$ $m_i^h \neq m^{K+1}$ but $m_j^h = m^{K+1}$, for all $j > i$, and for $K \geq h' > h$, $m_j^h = m^{K+1}$ for $j = 1, \ldots, n$, then $t_i = \delta$.

IMPLEMENTATION

Let ψ be the true profile of types. First note that if in each stage h, $h = 1, \ldots, K$, each of the n players announces the true vector, $m_i^h = \psi$, then $X(\psi)$ is implemented with probability of at least $1-\varepsilon$.

Next, we will show that truth telling is the unique subgame perfect equilibrium strategy in this game. Consider player n's choice in stage $K + 1$, and note that this is the last move in the game tree Γ. Announcing any $m_n^{K+1} \neq \psi_n$ will cost him at least $\varepsilon\Delta/n$ (in terms of reducing his expected utility) and will save him at most δ. Since $\delta < \varepsilon\Delta/n$, truth telling is the best move for player n at this stage. Having proved this for player n, we can move backwards and prove by induction that the same holds for every $i \neq n$ at stage $K + 1$. We have thus established that in any subgame perfect equilibrium, the profile ψ is obtained in stage $K + 1$.

Consider now player nth decision in stage K. Let $h' = \max\{h \mid m_n^h \neq \Psi\}$ if for some h, $m_n^h \neq \Psi$, otherwise, let $h' = 0$.

There are three cases to examine:

(i) for all $i, j \notin n \quad m_i^K = m_j^K$;

(ii) there exist $\quad i, j, h \notin n$ such that $\quad m_i^K \neq m_j^K \neq m_h^K \neq m_i^K$, and

(iii) there exists some $j \neq n$ such that for all $i, h \neq j$ and $i, h \neq n \quad m_i^K = m_h^K \neq m_j^K$.

In the first two cases (i) and (ii), player n's choice at stage K will not affect the social alternative chosen in this stage. Let us consider each of the three cases separately.

Case (*i*). If $h' = 0$, or $h' > 0$ but for some $h' < h \leq K$ $m_i^h \neq \psi$ for $i \neq n$, player n will be fined δ if at this stage $m_n^K \neq \psi$. If, instead, $m_n^K = \psi$, player n will not be fined. On the other hand, if $h' > 0$ and for all $h > h'$, $m_i^h = \psi$, then player n is fined by δ, regardless of his report. Consequently, he is just indifferent between $m_n^K = \psi$ and $m_n^K \neq \psi$.

Case (*ii*). Player n will be fined δ if he reports $m_n^K \neq \psi$; he will not be fined at all if he reports $m_n^K = \psi$.

Case (*iii*). In this case, n's choice may determine the alternative selected at this stage. Given what we know about stage $K + 1$, announcing $m_n^K \neq \psi$ will cost him δ, while his maximum potential gain is ξ/K. Since $\xi/K < \delta$ announcing the truth is the dominant move for player n in stage K.

We conclude, therefore, that at stage K there are only two possibilities for player n; if he is not the last one to lie along the history up to this point, he is strictly better off telling the truth, and otherwise, he is indifferent between telling the truth and lying.

We can now complete the proof by inducting on all $i \in N$, $i \neq n$ in stage K, and then on all stages h, $h < K$, using exactly the same arguments. Note that at the empty history (before the beginning of the game), player $i \in N$ has not lied, therefore, at stage $h = 1$, and given the continuation of the game, he is not indifferent between telling the truth and lying, and the only best response for him is to tell the truth.

Finally, we discuss how our mechanism can be modified to fit into the initial environment where monetary fines are not allowed. We shall only present the basic idea; for a complete proof the reader is referred to Abreu and Matsushima (1992). Notice first that in the mechanism above, the only payments made were from the agents to the planner and furthermore the size of these payments could be made arbitrarily small, since we only required that $\bar{t}$ be greater than zero. When monetary fines are not possible, the planner can punish the players by reducing the probability he assigned to the alternative “chosen” at stage $K+1$. That is, in each case in which some player i was fined by some amount t_i in the above mechanism, he will now be fined by reducing the probability the planner assigns to the alternative $f_i(m_i^{K+1})$, instead. It can be shown that the size of these punishments can be adjusted so as to be as effective as the monetary fines in the original mechanism.

REFERENCES

Abreu, D., and Matsushima, H. (1992). “Virtual Implementation in Iterative Undominated Strategies: Complete Information,” *Econometrica* **60**, 993–1108.

Abreu, D., and Sen, A. (1991). “Virtual Implementation in Nash Equilibrium,” *Econometrica* **59**, 997–1021.

Gibbard, A. (1973). “Manipulation of Voting Schemes: A General Result,” *Econometrica* **41**, 587–602.

Hurwicz, L. (1972). “On Informationally Decentralized Systems,” in *Decision and Organization* (R. Radner and C. B. McGuire, Eds.). Amsterdam: North-Holland.

Jackson, M. (1992). Implementation of Undominated Strategies: A Look at Bounded Mechanisms,” *Review of Economic Studies* **59**.

Maskin, E. (1977). “Nash Equilibrium and Welfare Optimality,” memo, MIT.

Moore, J. (1991). “Implementation, Contracts and Renegotiation in Environments with Complete Information,” in *Advances in Economic Theory*: *Proceedings of the Sixth World Congress of the Econometric Society* (J.-J. Laffont, Ed.). Cambridge: Cambridge Univ. Press.

Repullo, R. (1990). “Lecture Notes on Incentive Theory,” unpublished manuscript.

Satterthwaite, M. (1975). “Strategy-Proofness and Arrow’s Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions,” *J. Econ. Theory* **10**, 187–217.