Games and Economic Behavior **27,** 38–63 (1999)
Article ID game.1998.0671, available online at http://www.idealibrary.com on IDEAL®

Interactive Implementation*

Sandeep Baliga†

MEDS Department, Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208

and

Tomas Sjöström

Department of Economics, Harvard University, Cambridge, Massachusetts 02138

Received August 16, 1996

We suppose the principal not only designs a mechanism, but can participate as a player. The result is a Bayesian model where one player, the principal, has no information, and the remaining players have complete information. We find a necessary and sufficient condition for implementation. In contrast to the standard model, in the exchange economy many *cardinal* rules, such as the utilitarian social welfare function, satisfy this condition and hence can be implemented. Compared to the literature on Bayesian implementation, our mechanisms are rather simple. The idea is that the agents announce a state of the world, while the principal announces a strategy profile for the agents. *Journal of Economic Literature* Classification Numbers: C72, D71, D78. © 1999 Academic Press

1. INTRODUCTION

In the standard literature on implementation, the principal (or "planner") is the designer of a mechanism, but not a player. He designs the mechanism so that all its equilibria meet his objectives, but he plays no part in the mechanism he creates. In contrast, we study *interactive implementation* where the principal is also a player. Apart from its intrinsic interest, this

*The idea for this paper appeared in discussions with Tom Palfrey. Our debt to him is obvious. We are grateful to Rajiv Vohra, two anonymous referees, and seminar participants at Harvard University for helpful comments. Sandeep Baliga acknowledges the financial support of ESRC Grant Number R000221728.

†To whom correspondence should be addressed. Kellogg GSM (MEDS), 2001 Sheridan Road, Evanston, IL 60208-2009; E-mail: baliga@nwu.edu.

0899-8256/99 $30.00

theory expands the set of implementable social choice rules beyond previous results.

Consider the traditional model of implementation with $n \geq 3$ agents who all know the true state of the world, as surveyed by Moore (1992). The principal is uninformed about the state of the world. The standard assumption, which we maintain in this paper, is that the principal cannot control which equilibrium is played, so every equilibrium must yield the optimal outcome.[1] Also, every optimal outcome should be an equilibrium outcome. Suppose there are two possible "states of the world" ϕ and θ, and let f be the social choice rule. If f is implemented and $f(\theta) \neq f(\phi)$, then the set of equilibria must be different in the two states. Then (if the principal does not play) some agent must have different (ordinal) preferences over at least two feasible outcomes in the two states. This is the preference reversal condition. Even with powerful techniques such a virtual implementation (Abreu and Sen, 1991), some agent's preferences must reverse over at least two feasible lotteries. In other words, the state must be *payoff relevant* to the agents. In particular, in the traditional model it is impossible to implement cardinal rules such as the utilitarian rule. Subjecting some agent's utility function to a positive linear transformation changes the utilitarian optimum, without changing the agent's preferences over lotteries. Hence the preference reversal condition is violated and the utilitarian rule cannot be implemented, even virtually, if the planner does not play.

Why does it help to include the planner as a player? Because becoming a player gives him *equilibrium knowledge*. In a Bayesian–Nash equilibrium, each player has correct beliefs about the strategies of all the other players. By explicitly incorporating a set of possible actions for himself in the mechanism, the principal gives himself an opportunity to act on his equilibrium knowledge when the game is played. (This knowledge cannot be "built into" the mechanism at the design stage, because before the game exists the planner is not yet a player and cannot have equilibrium knowledge.) In our mechanism the informed agents announce the state of the world,

[1] If the principal could "select" an equilibrium, then he did not have to worry about the existence of multiple bad equilibria, and his problem would be trivial. The following mechanism would suffice. Each agent announces an outcome, and if at least $n - 1$ agents announce the same outcome a, the mechanism implements a (otherwise it implements some arbitrary outcome). If in each state θ every agent announces precisely that outcome which is optimal for the principal when the state is θ, then no agent i can achieve anything by a unilateral deviation. Thus, there is always a Nash equilibrium which is optimal for the principal. If he could "select" this equilibrium, this mechanism would solve the principal's problem. On the other hand, *every* outcome is a Nash equilibrium outcome of this mechanism. By contemplating this mechanism it becomes clear that the assumption that the principal can select an equilibrium is unpalatable, whether or not the principal is a player. Thus, we require that *every* equilibrium is optimal.

and the principal (who does not know the state) simultaneously announces *the agents' strategies*. The outcome function will basically try to implement whichever outcome is preferred by the principal, given the agents' reports about the state and the principal's report about the agents' strategies.[2]

Suppose there are two states ϕ and θ and $f(\theta) \neq f(\phi)$. Even if the agents have the same ordinal preferences in both states, f may be interactively implementable by our mechanism. In a separating equilibrium (where the agents send different messages in different states), the principal must get the optimal outcome in both states. Indeed he can get it by simply announcing the agents' strategies truthfully. The mechanism will "invert" the strategies, and pick the right outcome. Even if the agents always lie and say θ when the state is ϕ and vice versa, the planner "knows it" and can report this, so that the outcome is optimal for state ϕ if the agents say θ, and vice versa. Can there exist a pooling equilibrium, where the agents send the same message in both states? In such an equilibrium, the best outcome for the principal would be the outcome, say a, that maximizes his expected payoff given his prior beliefs about the likelihood of θ and ϕ. Suppose $a \neq f(\theta)$ and $a \neq f(\phi)$. Then the principal's best choice is to truthfully "tell" the mechanism that the equilibrium is pooling, and to announce a high integer. Then the outcome will be a. However, the mechanism is so constructed that if the principal says that the equilibrium is pooling, any agent can get his most preferred outcome by announcing an integer which is higher than anybody else's, including the principal. Such an integer game is in general incompatible with equilibrium. Thus, a pooling equilibrium will in general not exist. Since any separating equilibrium is optimal, this shows how f can be interactively implemented.

Notice that in the above example, we did not assume preference reversal for the agents. We did assume $a \neq f(\theta)$ and $a \neq f(\phi)$. If this assumption is *not* satisfied, the principal may have no incentive to report truthfully that the agents pool; hence the integer game will not be triggered and a bad equilibrium may, but need not, persist (we will give a precise necessary and sufficient condition). But intuitively, if we consider an economic environment, we would be surprised to find that a planner who does not know whether the state is ϕ or θ prefers exactly $f(\theta)$ (or exactly $f(\phi)$).

[2]It may be argued that the assumption that the principal has equilibrium knowledge is too strong. But we do not know any good argument for why principals should warrant a special exception from equilibrium analysis. In fact all the well-known "stories" can be used to justify equilibrium analysis in this model. For example, the players (including the principal) may meet before the game is played to discuss their plans, and the Bayesian–Nash equilibria are the self-enforcing agreements they could reach. Such pre-play communication may open up possibilities for collusion among the agents against the principal, but Baliga and Sjöström (1996) show that, at least in the exchange economy, collusion does not introduce any new restrictions apart from a condition of Pareto-optimality in the distribution of goods.

That is, we expect that the "compromise" alternative a satisfies $a \neq f(\theta)$ and $a \neq f(\phi)$, at least if the planner's preferences are sufficiently smooth. Certainly this will be the case for a utilitarian planner (although not for an egalitarian one, whose preferences are not smooth). This is discussed in Section 4. Perhaps the fact that f can be interactively implemented even though the state is not payoff relevant to the agents is not surprising, given that the state is payoff relevant *to the principal* if $f(\theta) \neq f(\phi)$, and the principal is a player in our model.

Throughout the paper, we assume the principal has irrevocably committed to the mechanism. If a bad outcome results, it is not subject to renegotiation among the principal and agents. It is, in fact, a standard assumption in the literature that the mechanism designer can credibly commit in this way. The principal's credibility problem is analyzed by Baliga *et al.* (1997) and Chakravorty *et al.* (1997).

Formally, our model is a special case of the Bayesian models developed by Jackson (1991), Palfrey (1992), Palfrey and Srivastava (1989) and Postlewaite and Schmeidler (1986). Palfrey (1992), Palfrey and Srivastava (1989), Baliga (1993), and Postlewaite and Schmeidler (1986) have pointed out that adding a completely uninformed player can change the set of implementable social choice rules. In our paper, the uninformed player is the principal himself (all other players have complete information). Of course, the principal is a very special player, as he is also the "mechanism designer" and his preferences coincide with the social choice rule. For general environments, there does not exist a necessary and sufficient condition for Bayesian implementation which can be translated to our interactive model. However, given the special structure of our problem, we find a necessary and sufficient condition which is closely related to the results on Nash implementation contained in Moore and Repullo (1990) and Sjöström (1991).

We believe interactive implementation is relevant for many applications. For example, in a contracting model there may exist an arbitrator or "judge." What is the judge's objective function? Ideal judges should not have any personal stake in the decision (Posner, 1993). Both traditional legal theories and "law and economics" theories suggest that the judge should favor outcomes that the contracting parties would have agreed on *ex ante*, if they could have signed an enforceable complete contract specifying the appropriate outcome in each state of the world (Schwartz, 1992). For example, the agents may have wanted an efficient outcome with an equitable distribution of the surplus in the various states. This complete contract would correspond to a "social welfare function" (map from states to outcomes). Posner (1993) argues that judges in fact get *utility* from following the legal theory, much like a spectator of a play derives utility from supporting the hero of the play. Such a judge is an "intervenor" in the sense of Hurwicz (1993): his utility function agrees with the *ex ante* objec-

tives of the contracting parties. Of course, the judge's problem is that he cannot directly observe the true state of the world, although the agents themselves know the state. The mechanism is the arbitration or court proceeding, and the social choice rule is the function from states to outcomes which corresponds to a "complete contract." It is easy to construct a buyer–seller model where an efficient outcome cannot be implemented in a standard (non-interactive) way, because *sunk costs* are not payoff relevant *ex post* when the mechanism is played (see Baliga and Sjöström, 1996). However, sunk costs can be payoff relevant to a judge, for reasons of fairness, risk-sharing, or because the parties *ex ante* would want the *ex post* outcome to depend on sunk costs and the judge is an "intervenor." The judge may then be able to interactively implement the efficient outcome.

2. DEFINITIONS

The set of *agents* is $N = \{1, 2, \ldots, n\}$, $n \geq 3$. The set of *players* is $N^* = N \cup \{0\}$ where player 0 is the principal. Let A be the set of outcomes and let Θ be the finite set of possible states of the world. Let $\Delta(\Theta)$ be the set of all probability distributions over Θ.

The payoff to player $i \in N^*$ if the outcome is $a \in A$ and the state is θ is

$$u_i(a, \theta).$$

The *lower contour set* for agent $i \in N$ at outcome a in state θ is the set

$$L_i(a, \theta) = \{b \in A: u_i(a, \theta) \geq u_i(b, \theta)\}.$$

The agents, but not the principal, know θ. In other respects the principal is treated symmetrically with the agents. The principal's state-dependent utility function is $u_0(a, \theta)$. The prior probabilities over states are given by the vector $\pi \in \Delta^0(\Theta) \equiv \{\pi \in \Delta(\Theta): \pi(\theta) > 0, \forall \theta \in \Theta\}$. If $T \subseteq \Theta$, then we define

$$\pi(T) \equiv \sum_{\theta \in T} \pi(\theta).$$

Let $|T|$ denote the number of elements in the set T.

A *social choice rule* is a correspondence from Θ to A. In our setting, it corresponds to the principal's most preferred outcomes when the state is θ:

$$F(\theta) \equiv \arg\max_{a \in A} u_0(a, \theta). \tag{1}$$

In the game we construct, it may happen that the agents pool (send the same message) in a set of states T. We would like to know what the principal's best outcome is following such a message. Applying Bayes' rule,

we define the principal's utility of getting outcome a conditional on the set T occurring as follows:

$$u_0(a, T) \equiv \sum_{\theta \in T} \frac{\pi(\theta)}{\pi(T)} u_0(a, \theta).$$

We assume $\arg\max_{a \in A} u_0(a, T) \neq \varnothing$ for all $T \subseteq \Theta$. It is then convenient to extend the definition of F to all subsets $T \subseteq \Theta$ as follows:

$$F(T) \equiv \arg\max_{a \in A} u_0(a, T). \tag{2}$$

If $a \in F(T)$ and $|T| > 1$, then a is a "best compromise" for the set T. From now on, we fix the principal's utility function $u_0(\cdot, \cdot)$ and this also fixes F, from (1) and (2). The function $u_0(\cdot, \cdot)$ is common knowledge and not subject to change.

If f is a single-valued function such that $f(T) \in F(T)$ for all $T \subseteq \Theta$, then f is a *selection* from F, and we write $f \in F$. Let $\mathcal{F}$ be the set of all selections from F.

Throughout this paper, we make a very weak assumption of unanimity. It states that if there exists a set of types T and an outcome a such that (a) whenever the state is in T, all agents agree that a is the best possible outcome, and (b) the principal agrees that a is the best outcome conditional on the state belonging to T, then a is in fact optimal for *each state in* T:

ASSUMPTION 1. Unanimity: *If there exists a set* $T \subseteq \Theta$ *such that for all* $j \in N$,

$$a \in \bigcap_{\theta \in T} \arg\max_{a \in A} u_j(a, \theta)$$

and

$$a \in F(T),$$

then

$$a \in \bigcap_{\theta \in T} F(\theta)$$

Notice that this is even weaker than the usual unanimity assumption, which states that a should be chosen whenever *all the agents* think a is the best outcome. The case where Assumption 1 does not hold can be handled in a way similar to that discussed by Sjöström (1991).

A mechanism (M, h) consists of a message space M_i for each player $i \in N^*$, and an outcome function $h : M \equiv \mathsf{3}_{i=0}^{n} M_i \to A$. The agents and the principal simultaneously send messages $m_i \in M_i$, $i = 0, 1, \ldots, n$. Then, the outcome is $h(m) \in A$. The outcome cannot be renegotiated: we assume the principal has irrevocably committed to the outcome function h.

A *message profile* is denoted $m = (m_0, m_1, \ldots, m_n) = (m_{-j}, m_j)$, where $m_{-j} \equiv (m_0, \ldots, m_{j-1}, m_{j+1}, \ldots, m_n)$. The *attainable set* for agent $i \in N$ at message profile m is

$$h(M_i, m_{-i}) \equiv \{b \in A: b = h(m_i', m_{-i}) \quad \text{for some } m_i' \in M_i\}.$$

A *strategy* for agent $i \in N$ is a function $\sigma_i: \Theta \to M_i$. We follow the standard procedure of ruling out mixed strategies. Let Σ_i denote agent i's strategy space, and let $\Sigma = \mathbin{3}_{i=1}^{n} \Sigma_i$. Define

$$\sigma \equiv (\sigma_1, \ldots, \sigma_n),$$
$$\sigma(\theta) \equiv (\sigma_1(\theta), \sigma_2(\theta), \ldots, \sigma_n(\theta)),$$

and

$$\sigma_{-j}(\theta) \equiv (\sigma_1(\theta), \ldots, \sigma_{j-1}(\theta), \sigma_{j+1}(\theta), \ldots, \sigma_n(\theta)).$$

A strategy for the principal is simply a message $m_0 \in M_0$. A *strategy profile* for all $n+1$ players is denoted (m_0, σ). The agents know the state of the world, but the principal is uninformed. Therefore, the agents' messages can depend on the state, but the principal's message must be independent of the state.

DEFINITION 1. A strategy profile (m_0^*, σ^*) is a (pure strategy) *Bayesian–Nash equilibrium* of (M, h) if

$$\sum_{\theta \in \Theta} \pi(\theta) u_0(h(m_0^*, \sigma^*(\theta)), \theta)$$
$$\geq \sum_{\theta \in \Theta} \pi(\theta) u_0(h(m_0, \sigma^*(\theta)), \theta) \quad \text{for all } m_0 \in M_0$$

and for all $\theta \in \Theta$ and all $j \in N$,

$$u_j(h(m_0^*, \sigma^*(\theta)), \theta) \geq u_j(h(m_0^*, \sigma_{-j}^*(\theta), m_j), \theta) \quad \text{for all } m_j \in M_j.$$

For a given mechanism (M, h) let $E(M, h)$ denote the set of (pure strategy) Bayesian–Nash equilibria. We will consider *full implementation*, which requires that (i) each equilibrium yields an optimal outcome in each state, and (ii) each selection from the social choice rule can be achieved in some equilibrium. Formally:

DEFINITION 2. The mechanism (M, h) *interactively implements the principal's optimum* if (i) if $(m_0^*, \sigma^*) \in E(M, h)$, then

$$h(m_0^*, \sigma^*(\theta)) \in F(\theta) \equiv \arg\max_{a \in A} u_0(a, \theta)$$

for all $\theta \in \Theta$, and (ii) for any selection $f \in F$, there exists $(m_0^*, \sigma^*) \in E(M, h)$ such that $h(m_0^*, \sigma^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

3. A NECESSARY AND SUFFICIENT CONDITION FOR IMPLEMENTATION

We follow Moore and Repullo (1990) and Sjöström (1991) and discuss the properties of the attainable sets of a mechanism which interactively implements the principal's optimum. We need a few definitions. If $a^* \in B_i \subseteq A$, define

$$C_i(B_i, a^*) \equiv \{\theta \in \Theta : B_i \subseteq L_i(a^*, \theta) \text{ and } A \subseteq L_j(a^*, \theta) \text{ for all } j \in N\backslash\{i\}\}.$$

DEFINITION 3. Let $(i, \bar{\theta}, \bar{a}) \in N \times \Theta \times A$ be such that $\bar{a} \in F(\bar{\theta})$. A set B_i is *acceptable for* $(i, \bar{\theta}, \bar{a})$ if $\bar{a} \in B_i \subseteq L_i(\bar{a}, \bar{\theta})$, and whenever $a^* \in B_i$ satisfies

$$C_i(B_i, a^*) \neq \varnothing$$

and

$$a^* \in F(T)$$

for some non-empty set $T \subseteq C_i(B_i, a^*)$, then

$$a^* \in \bigcap_{\theta \in T} F(\theta).$$

We will motivate this definition in a way similar to Moore and Repullo (1990) and Sjöström (1991). These authors considered Nash implementation, while our solution concept is Bayesian–Nash. A comparison is made easier if we focus on what happens in one state, or a subset of states, and imagine the agents tell the truth in all other states. Notice that the agents in our model know the true state, so from the agents' point of view, it is as if a Nash equilibrium is played in each state.

Suppose $\bar{a} \in F(\bar{\theta})$. Then, there should exist an equilibrium that produces $\bar{a}$ in state $\bar{\theta}$. Let B_i denote player i's attainable set at this equilibrium in state $\bar{\theta}$. Clearly $\bar{a} \in B_i \subseteq L_i(\bar{a}, \bar{\theta})$. Now suppose a^* is attainable for player i, say, $a^* = h(m) \in B_i$, and suppose $\varnothing \neq T \subseteq C_i(B_i, a^*)$. Then if m is played at any $\theta \in T$, no agent has an incentive to deviate. Indeed, all agents except i get their favorite outcome overall because $A \subseteq L_j(a^*, \theta)$ for all $j \in N\backslash\{i\}$, and agent i gets his best outcome in B_i because $B_i \subseteq L_i(a^*, \theta)$. At this point, Moore and Repullo (1990) and Sjöström (1991) conclude that m is a Nash equilibrium for each $\theta \in T$, but in our model *the principal* may still break the equilibrium by deviating to a message m_0' such that $h(m_0', m_{-0}) \neq a^*$. However, if $a^* \in F(T)$, then the principal, conditional on the state being in T, would actually want the outcome to be a^* and would have no incentive to deviate. Then, if $a^* \notin F(\theta)$ for some $\theta \in T$

implementation fails because there is an equilibrium where the outcome at state θ is not optimal. Hence, we must require

$$a^* \in \bigcap_{\theta \in T} F(\theta).$$

That is, the attainable set must be *acceptable.* Of course, this is just a necessary condition for the mechanism.

The attainable sets should be as big as possible to make it easier for agents to deviate and knock out "bad" equilibria (see Moore and Repullo, 1990, and Sjöström, 1991). A good candidate for attainable set is therefore *the union* of the acceptable sets. Formally, for all $(i, \bar{\theta}, \bar{a}) \in N \times \Theta \times A$ such that $\bar{a} \in F(\bar{\theta})$, let $B_i^*(\bar{a}, \bar{\theta})$ denote the union of all sets B_i that are acceptable for $(i, \bar{\theta}, \bar{a})$. We now verify that this set is itself acceptable.

LEMMA 1. *For all* $(i, \bar{\theta}, \bar{a}) \in N \times \Theta \times A$ *such that* $\bar{a} \in F(\bar{\theta})$, *if* $B_i^*(\bar{a}, \bar{\theta}) \neq \varnothing$, *then* $B_i^*(\bar{a}, \bar{\theta})$ *is acceptable for* $(i, \bar{\theta}, \bar{a})$.

Proof. Clearly $\bar{a} \in B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \bar{\theta})$. Suppose $a^* \in B_i^*(\bar{a}, \bar{\theta})$ satisfies $C_i(B_i^*(\bar{a}, \bar{\theta}), a^*) \neq \varnothing$. For $\theta' \in C_i(B_i^*(\bar{a}, \bar{\theta}), a^*)$,

$$B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(a^*, \theta') \tag{3}$$

and

$$A \subseteq L_j(a^*, \theta') \tag{4}$$

for all $j \in N \backslash \{i\}$. Suppose in addition

$$a^* \in F(T) \tag{5}$$

for some non-empty set $T \subseteq C_i(B_i^*(\bar{a}, \bar{\theta}), a^*)$. We need to show

$$a^* \in \bigcap_{\theta \in T} F(\theta). \tag{6}$$

Since $a^* \in B_i^*(\bar{a}, \bar{\theta})$, there exists a set $B_i \subseteq B_i^*(\bar{a}, \bar{\theta})$ which is acceptable for $(i, \bar{\theta}, \bar{a})$, and which satisfies $a^* \in B_i$. Then (3) implies, for all $\theta' \in C_i(B_i^*(\bar{a}, \bar{\theta}), a^*)$,

$$B_i \subseteq L_i(a^*, \theta') \tag{7}$$

and since (4) holds for all $j \in N \backslash \{i\}$, we have

$$T \subseteq C_i(B_i^*(\bar{a}, \bar{\theta}), a^*) \subseteq C_i(B_i, a^*). \tag{8}$$

But B_i is acceptable, so (6) follows. ■

An important step in our main characterization result is the following lemma. It shows that the sets just introduced are the "maximal" attainable sets.

LEMMA 2. *If* (M, h) *is a mechanism which interactively implements the principal's optimum, and* $(m_0^*, \sigma^*) \in E(M, h)$, *then for all* $\bar{\theta} \in \Theta$ *and all* $i \in N$,

$$h\big(M_i, (m_0^*, \sigma_{-i}^*(\bar{\theta}))\big) \subseteq B_i^*(h(m_0^*, \sigma^*(\bar{\theta})), \bar{\theta}) \neq \varnothing.$$

Proof. Suppose the principal's optimum is implemented and $(m_0^*, \sigma^*) \in E(M, h)$. Let

$$\bar{a} = h(m_0^*, \sigma^*(\bar{\theta})) \in F(\bar{\theta}).$$

Clearly, it is enough to show that $h(M_i, (m_0^*, \sigma_{-i}^*(\bar{\theta}))$ is acceptable for $(i, \bar{\theta}, \bar{a})$. In order to derive a contradiction, define $B_i \equiv h(M_i, (m_0^*, \sigma_{-i}^*(\bar{\theta}))$ and suppose B_i is not acceptable for $(i, \bar{\theta}, \bar{a})$. Notice that $\bar{a} \in B_i \subseteq L_i(\bar{a}, \bar{\theta})$.

Since B_i is not acceptable, there exists $a^* \in B_i$ such that

$$C_i(B_i, a^*) \neq \varnothing \tag{9}$$

and for some $T \subseteq C_i(B_i, a^*)$

$$a^* \in F(T) \tag{10}$$

but for some $\theta' \in T$,

$$a^* \notin F(\theta'). \tag{11}$$

Let $\bar{m}_i \in M_i$ be such that

$$h\big(m_0^*, \bar{m}_i, \sigma_{-i}^*(\bar{\theta})\big) = a^*. \tag{12}$$

Consider the strategy $\bar{\sigma}$ defined by

$$\bar{\sigma}_i(\theta) = \begin{cases} \sigma_i^*(\theta) & \text{if } \theta \notin T \\ \bar{m}_i & \text{if } \theta \in T \end{cases}$$

and for $j \in N \backslash \{i\}$,

$$\bar{\sigma}_j(\theta) = \begin{cases} \sigma_j^*(\theta) & \text{if } \theta \notin T \\ \sigma_j^*(\bar{\theta}) & \text{if } \theta \in T \end{cases}$$

Consider now the strategy profile $(m_0^*, \bar{\sigma})$. By (11), (12) and the construction of $\bar{\sigma}$, there exists $\theta' \in T$ such that

$$h\big(m_0^*, \bar{\sigma}(\theta')\big) = h\big(m_0^*, \bar{m}_i, \sigma_{-i}^*(\bar{\theta})\big) = a^* \notin \arg\max_{a \in A} u_0(a, \theta'). \tag{13}$$

To prove the Lemma it suffices to show that $(m_0^*, \bar{\sigma}) \in E(M, h)$, as this together with (13) contradicts the assumption that the principal's optimum is implemented.

To see that $(m_0^*, \bar{\sigma}) \in E(M, h)$, first notice that as $(m_0^*, \sigma^*) \in E(M, h)$ and $\bar{\sigma}(\theta) = \sigma^*(\theta)$ for all $\theta \notin T$, no agent can improve at such θ. If, on the other hand $\theta \in T$, then recall that

$$a^* = h\big(m_0^*, \bar{\sigma}(\theta)\big)$$

and

$$B_i \equiv h(M_i, (m_0^*, \sigma_{-i}^*(\bar{\theta})) = h(M_i, (m_0^*, \bar{\sigma}_{-i}(\theta)).$$

Since $T \subseteq C_i(B_i, a^*)$ we have

$$h(M_i, (m_0^*, \bar{\sigma}_{-i}(\theta)) \subseteq L_i(a^*, \theta)$$

and

$$A \subseteq L_j(a^*, \theta)$$

for all $j \in N \backslash \{i\}$, so again no agent can improve.

Consider, finally, the principal. If $\theta \notin T$, then

$$u_0\big(h(m_0^*, \bar{\sigma}(\theta)), \theta\big) = u_0\big(h(m_0^*, \sigma^*(\theta)), \theta\big) = \max_{a \in A} u_0(a, \theta) = F(\theta)$$

because $(m_0^*, \sigma^*) \in E(M, h)$ and the principal's optimum is supposed to be implemented. For all $\theta \in T$ we have

$$\bar{\sigma}(\theta) = (\bar{m}_i, \sigma_{-i}^*(\bar{\theta}))$$

and

$$h(m_0^*, \bar{\sigma}(\theta)) = a^* \in \arg\max_{a \in A} u_0(a, T).$$

Therefore, for any $m_0' \in M_0$,

$$\sum_{\theta \in \Theta} \pi(\theta) u_0\big(h(m_0^*, \bar{\sigma}(\theta)), \theta\big) \geq \sum_{\theta \in \Theta} \pi(\theta) u_0\big(h(m_0', \bar{\sigma}(\theta)), \theta\big)$$

which proves $(m_0^*, \bar{\sigma}) \in E(M, h)$. ∎

We now introduce our main definition.

DEFINITION 4. *Interactive Monotonicity*: $B_i^*(\bar{a}, \bar{\theta}) \neq \varnothing$ for all $(i, \bar{\theta}, \bar{a}) \in N \times \Theta \times A$ such that $\bar{a} \in F(\bar{\theta})$. Moreover, if $\bar{\theta} \in T \subseteq \Theta$ and

$$\bar{a} \in F(\bar{\theta}) \cap F(T)$$

and

$$B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta)$$

for each $i \in N$ and each $\theta \in T$, then

$$\bar{a} \in \bigcap_{\theta \in T} F(\theta).$$

To motivate this definition, recall Maskin's (1985) results on Nash implementation. If

$$\bar{a} \in F(\bar{\theta}),$$

then there must exist equilibrium messages $\bar{m}$ such that $h(\bar{m}) = \bar{a}$. In the "canonical" mechanism for Nash implementation, agent i's attainable set is his lower contour set $B_i \equiv L_i(\bar{a}, \bar{\theta})$. If T is a set such that at each state $\theta \in T$ each agent's preferences suffer a Maskin-monotonic transformation, i.e., if

$$L_i(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta) \tag{14}$$

for each i and each $\theta \in T$, then no agent will deviate from $\bar{m}$ at any state in T. Maskin monotonicity requires that if these conditions are satisfied, then $\bar{a} \in F(\theta)$ for all $\theta \in T$. This is a necessary condition for Nash implementation. (Notice that it is without loss of generality to assume $\bar{\theta} \in T$, because trivially $L_i(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \bar{\theta})$ and $\bar{a} \in F(\bar{\theta})$ by assumption. This simplifies our later argument a bit.)

In our case, the attainable set is $B_i = B_i^*(\bar{a}, \bar{\theta})$ rather than $L_i(\bar{a}, \bar{\theta})$, so (14) is replaced by $B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta)$. Of course, the attainable sets must be non-empty. Also, if $\bar{m}$ is sent in state T in the interactive model, it is not necessarily part of an equilibrium because the principal may deviate. However, if $\bar{a} \in F(T)$, then the principal does not want to deviate. Interactive monotonicity requires that if these conditions are satisfied, then $\bar{a} \in F(\theta)$ for all $\theta \in T$.

We will show that interactive monotonicity is necessary and sufficient for interactive implementation. The requirement that $B_i^*(\bar{a}, \bar{\theta}) \neq \varnothing$ for all $(i, \bar{\theta}, \bar{a}) \in N \times \Theta \times A$ such that $\bar{a} \in F(\bar{\theta})$ is very weak and is always satisfied in economic environments (see the next section). Suppose this condition holds. Then, interactive monotonicity is satisfied if $\bar{a} \in \bigcap_{\theta \in T} F(\theta)$ whenever $\bar{a} \in F(\bar{\theta}) \cap F(T)$ for some $\bar{\theta} \in T$. This is a kind of regularity condition on the planner's preferences, which does not imply preference reversal for the agents. But suppose this regularity condition is violated. Then, interactive monotonicity is violated if $B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta)$ for each $i \in N$ and each $\theta \in T$. Notice that $B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta)$ if there is no preference reversal for the agents between the states θ and $\bar{\theta}$. This fact will be used in the next section to construct examples of social choice rules that *cannot* (even) be interactively implemented.

We first establish that interactive monotonicity is a necessary condition for interactive implementation.

PROPOSITION 1. *Interactive monotonicity is a necessary condition for interactive implementation of the principal's optimum*.

Proof. Suppose the principal's optimum is implemented by a mechanism (M, h). If

$$\bar{a} \in F(\bar{\theta}),$$

then there exists $(m_0^*, \sigma^*) \in E(M, h)$ such that $h(m_0^*, \sigma^*(\bar{\theta})) = \bar{a}$. From Lemma 2 we know that $B_i^*(\bar{a}, \bar{\theta}) \neq \varnothing$. Suppose there is $T \subseteq \Theta$ such that $\bar{\theta} \in T$,

$$B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta)$$

for each $i \in N$ and each $\theta \in T$ and

$$\bar{a} \in F(T). \tag{15}$$

We will show that

$$\bar{a} \in \bigcap_{\theta \in T} F(\theta).$$

Consider the strategy $\bar{\sigma}$ defined by: for all $j \in N$,

$$\bar{\sigma}_j(\theta) = \begin{cases} \sigma_j^*(\theta) & \text{if } \theta \notin T \\ \sigma_j^*(\bar{\theta}) & \text{if } \theta \in T \end{cases}$$

Claim. $(m_0^*, \bar{\sigma}) \in E(M, h)$.

Proof. Since $(m_0^*, \sigma^*) \in E(M, h)$, clearly no agent can improve at any $\theta \notin T$. Suppose $\theta \in T$. Lemma 2 implies that for all $i \in N$,

$$h(M_i, (m_0^*, \bar{\sigma}_{-i}(\theta)) = h(M_i, (m_0^*, \sigma_{-i}^*(\bar{\theta})) \subseteq B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta)$$

so each agent is using a best response at such θ too.

Finally, consider the principal. If $\theta \notin T$, then $\sigma^*(\theta) = \bar{\sigma}(\theta)$ so

$$u_0\big(h(m_0^*, \sigma^*(\theta)), \theta\big) = u_0\big(h(m_0^*, \bar{\sigma}(\theta)), \theta\big) = \max_{a \in A} u_0(a, \theta)$$

because $(m_0^*, \sigma^*) \in E(M, h)$ and the principal's optimum is supposed to be implemented. But (15) implies that if $\theta \in T$, then

$$h(m_0^*, \bar{\sigma}(\theta)) = h(m_0^*, \sigma^*(\bar{\theta})) = \bar{a} \in \arg\max_{a \in A} u_0(a, T).$$

Therefore, for any $m_0' \in M_0$,

$$\sum_{\theta \in \Theta} \pi(\theta) u_0(h(m_0^*, \bar{\sigma}(\theta)), \theta) \geq \sum_{\theta \in \Theta} \pi(\theta) u_0(h(m_0', \bar{\sigma}(\theta)), \theta)$$

which proves the claim.

Since the principal's optimum is implemented, the claim implies that

$$h(m_0^*, \bar{\sigma}(\theta)) = \bar{a} \in F(\theta)$$

for all $\theta \in T$, which completes the proof of the Proposition. ■

We shall establish the converse to Proposition 1 by assuming interactive monotonicity holds and exhibiting a mechanism which implements the principal's optimum.

Consider the following mechanism. Each agent $i \in N$ sends a message consisting of a state θ_i, an outcome a_i, and a positive integer z_i. A generic message is denoted

$$m_i = (\theta_i, a_i, z_i) \in M_i \equiv \Theta \times A \times Z,$$

where $Z = \{1, 2, \ldots\}$. The principal announces a *strategy* for the agents, a selection from F, and an integer. That is, the principal sends a message $m_0 \in M_0 \equiv \Sigma \times \mathcal{F} \times Z$. Let $m_0 = (\sigma_{01}, \ldots, \sigma_{0n}, f_0, z_0) \in M_0$ be a generic message, where σ_{0i} is the principal's announcement of player i's strategy. Let

$$\Psi(m) = \{\theta: \sigma_{0j}(\theta) = m_j \text{ for all } j \in N\}.$$

In other words, if the agents send messages $m_{-0} = (m_1, \ldots, m_n)$, and the principal announces some strategy σ_{0j} for each agent j, then $\Psi(m)$ are the states that are consistent with the agents and the principal's announcements. Clearly, if each agent j uses strategy σ_j, and the principal correctly announces $\sigma_{0j} = \sigma_j$ for each j, then the principal's and agents' messages will always be consistent, i.e., $\Psi(m) \neq \varnothing$ for all m.

The outcome function h: $\mathsf{3}_{i=0}^{n} M_i \to A$ is defined as follows.[3]

Rule 1. Suppose (i) $\sigma_{0i}(\theta) = (\theta, \cdot, \cdot)$ for all $i \in N$ and all $\theta \in \Theta$; and (ii) there is $\theta^* \in \Theta$ such that $m_i = (\theta^*, \cdot, \cdot)$ for all $i \in N$. Then $h(m) = f_0(\theta^*)$.

Rule 2. Suppose (i) $\sigma_{0i}(\theta) = (\theta, \cdot, \cdot)$ for all $i \in N$ and all $\theta \in \Theta$; and (ii) there is $j \in N$ and $\theta^* \in \Theta$ such that $m_i = (\theta^*, \cdot, \cdot)$ for all $i \in N \backslash \{j\}$, and $m_j = (\theta_j, a_j, z_j)$ where $\theta_j \neq \theta^*$. Then, if $a_j \in B_j^*(f_0(\theta^*), \theta^*)$ set $h(m) = a_j$. Otherwise, $h(m) = f_0(\theta^*)$.

Rule 3. All other cases. If $z_0 > z_i$ for all $i \in N$ and $\Psi(m) \neq \varnothing$, then $h(m) = f_0(\Psi(m))$. If for some $j \in N$, $z_j > z_i$ for all $i \in N^* \backslash \{j\}$, then $h(m) = a_j$. In all other cases, $h(m)$ can be arbitrary.

Rule 1 states that if the principal reports that the agents always "tell the truth," and the agents all report state θ^*, then the outcome will be

[3]To write down the outcome function we need to know $\mathcal{F}$; hence we need to know the principal's true preferences. This is reasonable since the principal himself designs the mechanism, presumably already aware of his own preferences. Also, it is similar to Maskin's (1985) mechanism which is tailored for a specific social choice rule. An alternative would be to have a mechanism which does not depend *directly* on the principal's *true* preferences, but where the principal has to announce his "type" at the same time as all other messages are sent. For example, when the principal designs the mechanism, he might not know whether he will later become utilitarian or egalitarian. When the game is played, part of the principal's message is to announce "egalitarian" or "utilitarian." We have not investigated this type of implementation.

optimal for state θ^*. Rule 2 states that if agent j deviates from a consensus and "asks for" an outcome a_j which is in the appropriate attainable set for him, then outcome a_j is chosen. However, if a_j does not belong to the attainable set, then the deviation is disregarded. Rule 3 is an integer game. If the principal announces the highest integer, then if there is a set of states $\Psi(m) \neq \varnothing$ that are compatible with the principal and agents' reports, the outcome is optimal for the planner conditional on the true state belonging to $\Psi(m)$. But if some agent has announced the highest integer, then this agent gets whatever he asks for.

Notice that if the principal knows what strategies the agents use, then by announcing the true strategies he can obtain an outcome which is optimal given the information partition induced by these strategies. If he reports that the agents are not truthful, then he will trigger Rule 3 in which case he should also announce a high integer.

The principal also picks a selection from F. This gives a new way of looking at full implementation. When the principal designs the mechanism, he need not decide which of the utility maximizing outcomes should occur in a particular state. When the time comes to play the game, he can break the tie by announcing some particular single-valued selection. In the standard literature the agents themselves have to break ties by coordinating on a particular optimal outcome. In our mechanism, there is no such coordination problem since the agents never act as tie-breakers.

LEMMA 3. *For any selection $f \in F$ there exists a "truth-telling" Bayesian–Nash equilibrium (m_0^*, σ^*) where $h(m_0^*, \sigma^*(\theta)) = f(\theta)$ for all θ.*

Proof. Let $m_0^* = (\sigma_1^*, \ldots, \sigma_n^*, f, \cdot)$ and $\sigma_i^*(\theta) = (\theta, \cdot, \cdot)$ for all $i \in N$ and all $\theta \in \Theta$. Then $h(m_0^*, \sigma(\theta)) = f(\theta)$ for all θ. Suppose the true state is θ. Agent $i \in N$ can only attain the outcome $a' \neq f(\theta)$ by a unilateral deviation if $a' \in B_i^*(f(\theta), \theta) \subseteq L_i(f(\theta), \theta)$, by Rule 2, but in this case agent i would not be made better off. The principal cannot improve, since by definition the outcome $f(\theta) \in F(\theta)$ maximizes his utility. ∎

LEMMA 4. *Suppose interactive monotonicity holds. If $(m_0^*, \sigma^*) \in E(M, h)$, then*

$$h(m_0^*, \sigma^*(\theta)) \in F(\theta)$$

for all $\theta \in \Theta$.

Proof. Let $(m_0^*, \sigma^*) \in E(M, h)$, where $m_0^* = (\sigma_{01}^*, \ldots, \sigma_{0n}^*, f_0^*, \cdot)$. Let $\mathcal{T} = \{T_1, \ldots, T_J\}$ be the (unique) partitioning of Θ satisfying: $\sigma^*(\theta) = \sigma^*(\theta')$ whenever $\theta, \theta' \in T_k \in \mathcal{T}$, and $\sigma^*(\theta) \neq \sigma^*(\theta')$ if $\theta \in T_k$, $\theta' \in T_l$, $k \neq l$. Abusing notation, we write $\sigma^*(T_k) = \sigma^*(\theta)$ if $\theta \in T_k$. Without loss of generality, suppose Rule 1 applies to $(m_0^*, \sigma^*(T_k))$ if $1 \leq k \leq K$, Rule 2

applies to $(m_0^*, \sigma^*(T_k))$ if $K < k \leq K'$, and Rule 3 applies to $(m_0^*, \sigma^*(T_k))$ if $K' < k \leq J$.

In Bayesian–Nash equilibrium, the principal knows σ^* and picks a message m_0 to maximize his expected payoff:

$$\sum_{k=1}^{J} \pi(T_k) u_0\big(h(m_0, \sigma^*(T_k)), T_k\big).$$

If the principal announces the agents' true strategies ($\sigma_{0i} = \sigma_i^*$ for all $i \in N$), picks some suitable f_0 and a sufficiently high integer, then he can guarantee himself the outcome

$$h(m_0, \sigma^*(T_k)) = f_0(T_k) \in \arg\max_{a \in A} u_0(a, T_k)$$

for each T_k so that his expected payoff is maximal:

$$\sum_{k=1}^{J} \pi(T_k) u_0\big(h(m_0, \sigma^*(T_k)), T_k\big) = \sum_{k=1}^{J} \pi(T_k) \times \max_{a \in A} u_0(a, T_k).$$

Therefore, since (m_0^*, σ^*) is a Bayesian–Nash equilibrium,

$$h(m_0^*, \sigma^*(T_k)) \in \arg\max_{a \in A} u_0(a, T_k) \text{ for all } T_k \in \mathcal{T}. \tag{16}$$

We need to show that for all T_k and all $\theta \in T_k$,

$$h(m_0^*, \sigma^*(T_k)) \in F(\theta). \tag{17}$$

Consider the state $\theta \in T_k$. First suppose $K < k \leq K'$, so Rule 2 applies to $(m_0^*, \sigma^*(T_k))$. If j is the agent with $\theta_j \neq \theta^* = \theta_i$ for all $i \in N \backslash \{j\}$, then by definition of Rule 2, as agent j is using a best response at state θ we must have

$$B_j^*(f_0^*(\theta^*), \theta^*) \subseteq L_j\big(h(m_0^*, \sigma^*(T_k)), \theta\big)$$

and by definition of Rule 3,

$$A \subseteq L_i\big(h(m_0^*, \sigma^*(T_k)), \theta\big)$$

for all $i \in N \backslash \{j\}$ (because these agents can trigger the integer game) so that

$$T_k \subseteq C_j\big(B_j^*(f_0^*(\theta^*), \theta^*), h(m_0^*, \sigma^*(T_k))\big).$$

We already know that

$$h(m_0^*, \sigma^*(T_k)) \in F(T_k).$$

Since $B_j^*(f_0^*(\theta^*), \theta^*)$ is acceptable by Lemma 1,

$$h(m_0^*, \sigma^*(T_k)) \in \bigcap_{\theta \in T_k} F(\theta)$$

so that (17) holds.

Next, suppose $K' < k \leq J$, so Rule 3 applies to $(m_0^*, \sigma^*(T_k))$. If the true state is $\theta \in T_k$, each agent can get his most preferred outcome by announcing a high integer, and the same is true for the principal. Since $(m_0^*, \sigma^*) \in E(M, h)$,

$$h(m_0^*, \sigma^*(T_k)) \in \arg\max_{a \in A} u_i(a, \theta) \tag{18}$$

for all $i \in N$ and all $\theta \in T_k$, and

$$h(m_0^*, \sigma^*(T_k)) \in F(T_k).$$

By unanimity, (17) holds.

Finally, suppose $1 \leq k \leq K$, so Rule 1 applies to $(m_0^*, \sigma^*(T_k))$. In order to obtain a contradiction, suppose there is $\theta^* \in T_k$ such that (17) does not hold, i.e.,

$$h(m_0^*, \sigma^*(T_k)) \equiv \bar{a} \notin F(\theta^*). \tag{19}$$

From (16) and (19) we deduce that $T_k \neq \{\theta^*\}$. But by definition of Rule 1 we have $\bar{a} = f_0^*(\bar{\theta}) \in F(\bar{\theta})$ for some $\bar{\theta} \in \Theta$. Let $T \equiv T_k \cup \{\bar{\theta}\}$. Then $\theta^*, \bar{\theta} \in T$ and

$$\bar{a} \in \arg\max_{a \in A} u_0(a, \bar{\theta}) \cap \arg\max_{a \in A} u_0(a, T_k) \subseteq \arg\max_{a \in A} u_0(a, T) \tag{20}$$

but

$$\bar{a} \notin \arg\max_{a \in A} u_0(a, \theta^*) \tag{21}$$

from (19). Since $(m_0^*, \sigma^*) \in E(M, h)$, by definition of Rule 2 we must have $B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta)$ for each $i \in N$ and each $\theta \in T_k$, or else some agent would deviate at such θ. Since $B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \bar{\theta})$ by construction, in fact

$$B_i^*(\bar{a}, \bar{\theta}) \subseteq L_i(\bar{a}, \theta) \tag{22}$$

for all $i \in N$ and all $\theta \in T$. However, this together with (20) and (21) contradicts the definition of interactive monotonicity. Hence, (19) cannot hold. Thus, (17) holds. ■

We have established:

THEOREM 1. *The principal can interactively implement his optimum if and only if interactive monotonicity holds.*

The condition of interactive monotonicity is relatively straightforward to check, as the sets $B_i^*(a, \theta)$ can be constructed by a method similar to the one described in Sjöström (1991). Start with the lower contour set $L_i(a, \theta)$, and check if $L_i(a, \theta)$ is "acceptable." If yes, then $B_i^*(a, \theta) = L_i(a, \theta)$. If not, there is some "bad outcome" a^* in $L_i(a, \theta)$ which causes $L_i(a, \theta)$ to violate the definition of acceptability. These are points that could lead to undesirable equilibria, if they were included in the attainable sets. Remove all these bad outcomes from $L_i(a, \theta)$. Call the new set B^1. It may now happen that some outcomes that were not "bad" in $L_i(a, \theta)$ have become bad in B^1. Remove them, etc. It can be shown that when there are no more bad points to remove, the final set is $B_i^*(a, \theta)$. Rather than going through the details for the general case, we will discuss interactive monotonicity in the exchange economy in the next section.

We make one final remark. We will show in the next section that it is possible to interactively implement a class of *cardinal* social welfare functions (these are, of course, not Nash implementable). We now show by example that it is also possible to interactively implement *non-cardinal* social choice rules which are not Nash implementable. Suppose there are three agents with identical preferences over four outcomes $\{a, b, c, d\}$, and $\Theta = \{\alpha, \beta\}$. In state α their preferences are $d \succ b \succ a \succ c$ and in state β they are $d \succ a \succ b \succ c$. The social choice rule is "pick the third-ranked outcome" in each state, i.e., $F(\alpha) = a$ and $F(\beta) = b$. This F is ordinal but is not Maskin monotonic and is therefore not Nash implementable. Suppose $F(\{\alpha, \beta\}) = c$ (this can be easily justified by a utility function and prior for the planner). Then interactive monotonicity holds so this social choice criterion is interactively implementable.

4. SOCIAL WELFARE FUNCTIONS IN ECONOMIC ENVIRONMENTS

Suppose the principal wants to allocate m divisible goods among the agents. Let a_{ik} denote agent i's consumption of good k, and let $a_i = (a_{i1}, \ldots, a_{im})$ denote agent i's consumption bundle. Let ω_k denote the amount available of the kth good. The feasible set is

$$A = \left\{ a \in \mathbf{R}_+^{nm} : a_{ik} \geq 0, \sum_{i=1}^{n} a_{ik} \leq \omega_k \right\}. \tag{23}$$

Each agent has selfish preferences: if $a_i' = a_i$ and $b_i' = b_i$, then

$$u_i(b, \theta) > u_i(a, \theta) \quad \Leftrightarrow \quad u_i(b', \theta) > u_i(a', \theta). \tag{24}$$

By (24), we can abuse notation and write for all $i \in N$

$$u_i(a, \theta) = u_i(a_i, \theta).$$

For each $i \in N$, and each $\theta \in \Theta$, $u_i(a_i, \theta)$ is differentiable and strictly concave in a_i, and strictly monotone:

$$\frac{\partial u_i}{\partial a_{ik}} > 0, \ \forall k.$$

In this environment, $n-1$ agents will never agree on a best possible outcome, so $C_i(B_i, a) = \varnothing$ for any B_i and a, and therefore

$$B_i^*(a, \theta) = L_i(a, \theta)$$

for all i, a and θ, which makes interactive monotonicity easy to check. Interactive monotonicity now states that if there is $T \subseteq \Theta$ such that outcome $\bar{a}$ is optimal in state $\bar{\theta} \in T$, and it is also the best compromise in set T, and if the agents' θ preferences for each $\theta \in T$ are Maskin-monotonic transformations of the $\bar{\theta}$ preferences at outcome $\bar{a}$, then $\bar{a}$ is optimal at each state $\theta \in T$.

Interactive monotonicity is clearly much weaker than the standard condition of Maskin monotonicity, which is necessary for Nash implementation (Maskin, 1985). In fact, as long as the "compromise" alternatives in $F(T)$ for $|T| > 1$ do not coincide with $F(\bar{\theta})$ for a particular $\bar{\theta} \in T$, interactive monotonicity is automatically satisfied. (Obviously, if $|T| = 1$, then interactive monotonicity cannot be violated with this T). This suggests that if there is sufficient smoothness, interactive monotonicity will hold, because "generically" optimal compromises would not be optimal if the state was actually known.

Let $W(u_1, \ldots, u_n)$ be a social welfare function. A social welfare maximizing principal has the utility function

$$u_0(a, \theta) = W\big[u_1(a_1, \theta), \ldots, u_n(a_n, \theta)\big]. \tag{25}$$

For the sake of clarity we make explicit that, conditional on a set T occurring ($|T| \geq 2$), the principal's preferences depend on the prior π. So, instead of $u_0(a, T)$, write

$$u_0(a, T, \pi) \equiv \sum_{\theta \in T} \frac{\pi(\theta)}{\pi(T)} W\big[u_1(a_1, \theta), \ldots, u_n(a_n, \theta)\big].$$

Fix a domain Θ. For this domain a property P holds *generically in the prior distribution* if there exists an open and dense set $X \subseteq \Delta^0(\Theta)$ such that property P is true whenever the prior beliefs belong to the set X.

THEOREM 2. *Suppose the principal's utility function is given by* (25), *where W is strictly increasing in each coordinate u_i, twice continuously differentiable, and concave. Suppose that for each $\theta \in \Theta$, the social welfare maximizing outcome is interior* ($a_{ik} > 0, \forall i, k$). *Then generically in the prior beliefs, the principal can interactively implement his optimum.*

Proof. We need to show that interactive monotonicity holds for generic $\pi \in \Delta^0(\Theta)$. Under our assumptions, a necessary and sufficient condition for a to be a social welfare maximum is that it satisfies the first order conditions

$$W_i \frac{\partial u_i(a_i, \theta)}{\partial a_{ik}} - W_n \frac{\partial u_n\left(\omega - \sum_{j=1}^{n-1} a_j, \theta\right)}{\partial a_{nk}} = 0, \quad i = 1, \ldots, n-1 \tag{26}$$

for all $k \in \{1, \ldots, m\}$, where

$$W_i \equiv \frac{\partial W}{\partial u_i}\left[u_1(a_1, \theta), \ldots, u_n(\omega - \sum_{j=1}^{n-1} a_j, \theta)\right].$$

Now fix any $T \subseteq \Theta$ with $|T| \geq 2$. There are two possibilities.

Case 1. There exists $a^* \in A$ such that for all $\theta \in T$, and all $k \in \{1, \ldots, m\}$,

$$W_i \frac{\partial u_i(a_i^*, \theta)}{\partial a_{ik}} - W_n \frac{\partial u_n\left(\omega - \sum_{j=1}^{n-1} a_j^*, \theta\right)}{\partial a_{nk}} = 0, \quad i = 1, \ldots, n-1.$$

For this a^*, (26) holds for all $\theta \in T$, so that

$$a^* = \bigcap_{\theta \in T} \arg\max_{a \in A} W\left[u_1(a_1, \theta), \ldots, u_n(a_n, \theta)\right].$$

Clearly, interactive monotonicity cannot be violated with this T for any prior beliefs $\pi \in \Delta^0(\Theta)$.

Case 2. For all $a^* \in A$, there is $\theta \in T$ and $k \in \{1, \ldots, m\}$ and $i \in N$ such that

$$W_i \frac{\partial u_i(a_i^*, \theta)}{\partial a_{ik}} - W_n \frac{\partial u_n\left(\omega - \sum_{j=1}^{n-1} a_j^*, \theta\right)}{\partial a_{nk}} \neq 0. \tag{27}$$

To show that (generically) interactive monotonicity cannot be violated with this T, it suffices to show that generically in the prior probabilities, for all $\theta \in T$

$$\arg\max_{a \in A} u_0(a, T, \pi) \neq \arg\max_{a \in A} u_0(a, \theta). \tag{28}$$

Let

$$a^* = \arg\max_{a \in A} u_0(a, T, \pi).$$

If a^* is not an interior point of A, then (28) holds by the interiority assumption, so we can suppose a^* is interior. Then for all $k \in \{1, \ldots, m\}$,

$$\sum_{\theta \in T} \pi(\theta)\left[W_i \frac{\partial u_i(a_i^*, \theta)}{\partial a_{ik}} - W_n \frac{\partial u_n(\omega - \sum_{j=1}^{n-1} a_j^*, \theta)}{\partial a_{nk}}\right] = 0, \quad i = 1, \ldots, n-1. \tag{29}$$

By (29) and (27) there is $i \in N$ and $k \in \{1, \ldots, m\}$ and $\theta', \theta'' \in T$ such that

$$W_i \frac{\partial u_i(a_i^*, \theta')}{\partial a_{ik}} - W_n \frac{\partial u_n\left(\omega - \sum_{j=1}^{n-1} a_j^*, \theta'\right)}{\partial a_{nk}} > 0 \tag{30}$$

and

$$W_i \frac{\partial u_i(a_i^*, \theta'')}{\partial a_{ik}} - W_n \frac{\partial u_n\left(\omega - \sum_{j=1}^{n-1} a_j^*, \theta''\right)}{\partial a_{nk}} < 0. \tag{31}$$

Define

$$X_T \equiv \left\{\pi \in \Delta^0(\Theta)\text{: } \arg\max u_0(a, T, \pi) \neq \arg\max u_0(a, \theta) \text{ for all } \theta \in T\right\}.$$

If $\pi \in X_T$, then interactive monotonicity cannot be violated with this T and these priors. By a standard argument, $\arg\max u_0(a, T, \pi)$ depends continuously on π, so X_T is open in $\Delta^0(\Theta)$. If $\pi \notin X_T$ then define π' by: $\pi'(\theta') = \pi(\theta') + \varepsilon$ and $\pi'(\theta'') = \pi(\theta'') - \varepsilon$, where θ' and θ'' satisfy (30) and (31), and let $\pi'(\theta) = \pi(\theta)$ for $\theta \neq \theta', \theta''$. For arbitrarily small $\varepsilon \neq 0$, $\pi' \in X_T$ because (29) will not hold with the prior π'. Therefore, X_T is dense in $\Delta^0(\Theta)$. Finally, set

$$X = \bigcap_{T \subseteq \Theta} X_T \subseteq \Delta^0(\Theta)$$

to get the open and dense set for which interactive monotonicity holds. ■

Theorem 2 implies that the utilitarian criterion can be interactively implemented for almost all priors, as long as the interiority assumption is satisfied. (The interiority assumption will be satisfied if, for example, utility functions have the Cobb–Douglas form.) The utilitarian optimum may differ in two states θ and θ' even if there is no preference reversal for any agent, so powerful non-interactive techniques such as virtual implementation fail for this rule. More generally, weighted *CES* social welfare functions of the form

$$W[u_1, \ldots, u_n] = \left(r_1 u_1^\rho + r_2 u_2^\rho + \cdots + r_n u_n^\rho\right)^{1/\rho},$$

where $r_i > 0$, $\forall i$, $\sum_{i=1}^n r_i = 1$ and $-\infty < \rho \leq 1$, can be interactively implemented, as long as the interiority assumption is satisfied. (The egalitarian criterion can be obtained as a limit as $\rho \to -\infty$, but then smoothness is lost and as shown below interactive implementation fails in general.)

We now argue that Theorem 2 is actually tight in the sense that implementation may not be possible if either (a) the priors are non-generic, or (b) the hypotheses of the theorem are dropped. Consider first the following example.

EXAMPLE 1. There are three states, $\Theta = \{\theta', \theta'', \theta'''\}$ and three agents with continuous preferences. Agents 2 and 3's preferences do not depend on the state. Thus, for $i = 2, 3$ we can write $u_i(a_i, \theta) = u_i(a_i)$. Agent 1's (cardinal) preferences depend on the state. Specifically, there exists a function $u_1(a_1)$ such that

$$u_1(a_1, \theta') = u_1(a_1),$$
$$u_1(a_1, \theta'') = 2u_1(a_1),$$
$$u_1(a_1, \theta''') = 3u_1(a_1).$$

Suppose $\pi(\theta') = \pi(\theta'') = \pi(\theta''') = 1/3$. The principal is utilitarian. Suppose the utility functions are such that the utilitarian optimum is interior, $a_{ik} > 0$ for all i, k. Consider

$$\begin{aligned} u_0(a, \Theta, \pi) &= \sum_{\theta \in T} \pi(\theta) \left[\sum_{i=1}^{3} u_i(a_i, \theta) \right] \\ &= \frac{u_1(a_1) + 2u_1(a_1) + 3u_1(a_1)}{3} + u_2(a_2) + u_3(a_3) \\ &= 2u_1(a_1) + u_2(a_2) + u_3(a_3) \ = u_0(a, \theta''). \end{aligned} \tag{32}$$

Therefore, there exists a^* such that

$$a^* \in \arg\max_{a \in A} u_0(a, \Theta, \pi) = \arg\max_{a \in A} u_0(a, \theta'')$$

and since the utilitarian optimum is clearly different in the three states,

$$a^* \notin \bigcap_{\theta \in \Theta} \arg\max_{a \in A} u_0(a, \theta).$$

Moreover, there is no preference reversal for any agent as ordinal utilities are always the same. Hence, interactive monotonicity is not satisfied (cf. the discussion preceding Proposition 1), and the principal cannot interactively implement his optimum.

This example does not contradict Theorem 2 because if we perturb probabilities away from (1/3, 1/3, 1/3), maximizing $u_0(a, \Theta, \pi)$ will not be the same as maximizing $u_0(a, \theta)$ for some $\theta \in \Theta$. Then interactive monotonicity will be satisfied.

If the social welfare function W does not satisfy the hypotheses of Theorem 2, then failure of implementation can happen even for generic priors. In particular, if W is not differentiable then the principal's best outcome for some specific state may very well also be his best compromise for some set. Consider the egalitarian social welfare function:

$$W(u_1, \ldots, u_n) = \min_{i \in N} u_i.$$

For the egalitarian principal,

$$u_0(a, \theta) = \min_{i \in N} u_i(a_i, \theta).$$

Clearly, once we rule out corner solutions, then

$$\arg\max u_0(a, \theta) \in \{a \in A: u_1(a_1, \theta) = u_2(a_2, \theta) = \cdots = u_n(a_n, \theta)\}. \quad (33)$$

Due to the non-differentiability, the egalitarian solution cannot in general be interactively implemented. Since the argument concerns cardinal utilities, let us assume that it is possible for at least one agent to suffer a scaling of the utility function. Moreover, to simplify calculations we suppose the domain includes some Cobb–Douglas utility function. Formally:

ASSUMPTION CD. *There exist* $\theta, \theta' \in \Theta$ *and* $j \in N$ *such that* (i)–(iii) *hold*: (i) *for all* $i \in N$,

$$u_i(\cdot, \theta) = \prod_{k=1}^{m} a_{ik}^{\alpha_k}, \quad \textit{where } 0 < \alpha_k < 1 \textit{ for all } k;$$

(ii) $u_i(\cdot, \theta') = u_i(\cdot, \theta)$ *if* $i \in N \backslash \{j\}$; *and* (iii) $u_j(\cdot, \theta') = \beta u_j(\cdot, \theta)$ *for some* $\beta > 1$.

PROPOSITION 2. *If Assumption CD holds, and if either*

$$\frac{\pi(\theta)}{\pi(\theta')}(n-1) < 1/\beta$$

or

$$\frac{\pi(\theta)}{\pi(\theta')}(n-1) > \beta,$$

then the egalitarian principal cannot interactively implement his optimum.

Proof. Let the two states θ and θ' be as defined in Assumption CD. Without loss of generality suppose $j = 1$, and also for simplicity

$$\sum_{k=1}^{m} \alpha_k = 1.$$

Consider

$$u_0(a, \{\theta, \theta'\}, \pi) \equiv \pi(\theta) u_0(a, \theta) + \pi(\theta') u_0(a, \theta')$$

Suppose $\hat{a} = \arg\max u_0(a, \{\theta, \theta'\}, \pi)$.

Claim. Either $\hat{a} = \arg\max u_0(a, \theta)$ or $\hat{a} = \arg\max u_0(a, \theta')$.

Proof of claim. First, note that an argument similar to that which leads to (33) establishes that, since players $i, j \in N\backslash\{1\}$ have identical utility functions in states θ and θ', we must have $\hat{a}_i = \hat{a}_j$ for all $i, j \in N\backslash\{1\}$. Thus, we may consider only agent 1 and some other agent $j \neq 1$, say, $j = 2$. In fact,

$$\begin{aligned} u_0(\hat{a}, \{\theta, \theta'\}, \pi) &= \max_{a \in A} u_0(a, \{\theta, \theta'\}, \pi) \\ &= \max_{a \in A} \{\pi(\theta) u_0(a, \theta) + \pi(\theta') u_0(a, \theta')\} \\ &= \max_{a \in A} \left\{ \pi(\theta) \min \left\{ \prod_{k=1}^{m} a_{1k}^{\alpha_k}, \prod_{k=1}^{m} a_{2k}^{\alpha_k} \right\} \right. \\ &\qquad \left. + \pi(\theta') \min \left\{ \beta \prod_{k=1}^{m} a_{1k}^{\alpha_k}, \prod_{k=1}^{m} a_{2k}^{\alpha_k} \right\} \right\}. \end{aligned}$$

It is obvious that $\hat{a}$ must satisfy

$$\prod_{k=1}^{m} \hat{a}_{1k}^{\alpha_k} \leq \prod_{k=1}^{m} \hat{a}_{2k}^{\alpha_k} \leq \beta \prod_{k=1}^{m} \hat{a}_{1k}^{\alpha_k}, \tag{34}$$

so that in fact

$$u_0(\hat{a}, \{\theta, \theta'\}, \pi) = \pi(\theta) \prod_{k=1}^{m} \hat{a}_{1k}^{\alpha_k} + \pi(\theta') \prod_{k=1}^{m} \hat{a}_{2k}^{\alpha_k}. \tag{35}$$

If the first weak inequality in (34) holds with equality, it must be that $\hat{a}_1 = \hat{a}_2 = \hat{a}_j$ for all $j > 2$, so

$$\hat{a} = \arg\max_{a \in A} u_0(a, \theta).$$

Similarly, if the second weak inequality holds with equality, then

$$\hat{a} = \arg\max_{a \in A} u_0(a, \theta').$$

Thus, if our claim is incorrect, then

$$\prod_{k=1}^{m} \hat{a}_{1k}^{\alpha_k} < \prod_{k=1}^{m} \hat{a}_{2k}^{\alpha_k} < \beta \prod_{k=1}^{m} \hat{a}_{1k}^{\alpha_k}. \tag{36}$$

Then the following program has an interior solution: choose $a \in A$ to maximize

$$\pi(\theta) \prod_{k=1}^{m} a_{1k}^{\alpha_k} + \pi(\theta') \prod_{k=1}^{m} a_{2k}^{\alpha_k} \tag{37}$$

subject to

$$\prod_{k=1}^{m} a_{1k}^{\alpha_k} \leq \prod_{k=1}^{m} a_{2k}^{\alpha_k} \leq \beta \prod_{k=1}^{m} a_{1k}^{\alpha_k} \tag{38}$$

and

$$a_j = a_2, \quad \forall j > 2.$$

By interior solution, we mean that both inequalities in (38) are strict. Thus, in the maximization program we can neglect (38), set

$$a_1 = \omega - \sum_{j=2}^{n} a_j = \omega - (n-1)a_2$$

and choose a_2 to maximize

$$\pi(\theta) \prod_{k=1}^{m} (\omega_k - (n-1)a_{2k})^{\alpha_k} + \pi(\theta') \prod_{k=1}^{m} a_{2k}^{\alpha_k}.$$

The first order necessary conditions are: for $k = 1, \ldots, m$,

$$-\pi(\theta)(n-1)\frac{\alpha_k}{a_{1k}} u_1(a_1, \theta) + \pi(\theta')\frac{\alpha_k}{a_{2k}} u_2(a_2, \theta) = 0.$$

Since the inequalities in (38) are assumed to hold, this implies

$$1 \leq \frac{\pi(\theta)}{\pi(\theta')}(n-1)\frac{a_{2k}}{a_{1k}} = \frac{u_2(a_2, \theta)}{u_1(a_1, \theta)} \leq \beta.$$

Thus, if we suppose

$$\frac{\pi(\theta)}{\pi(\theta')}(n-1)\beta < 1, \tag{39}$$

then

$$\frac{a_{2k}}{a_{1k}} > \beta \quad \text{for } k = 1, \ldots, m,$$

which implies

$$\beta \prod_{k=1}^{m} a_{1k}^{\alpha_k} = \prod_{k=1}^{m} (\beta a_{1k})^{\alpha_k} < \prod_{k=1}^{m} a_{2k}^{\alpha_k},$$

contradicting (38). A similar contradiction obtains if we suppose

$$\frac{\pi(\theta)}{\pi(\theta')}(n-1) > \beta. \tag{40}$$

Since we are assuming that either (39) or (40) holds, this proves the claim.

Suppose actually $\hat{a} = \arg\max u_0(a, \theta)$. Clearly, $\hat{a} \notin \arg\max u_0(a, \theta')$. Since there is no preference reversal for any agent, interactive monotonicity does not hold. Similarly, if $\hat{a} = \arg\max u_0(a, \theta')$, then $\hat{a} \notin \arg\max u_0(a, \theta)$ and again interactive monotonicity fails. ■

REFERENCES

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

Baliga, S. (1993). "Uncertainty in Implementation and Bargaining," Ph.D. thesis, Harvard.

Baliga, S., Corchon, L., and Sjöström, T. (1997). "The Theory of Implementation when the Planner is a Player," *J. Econ. Theory* **77**, 15–33.

Baliga, S., and Sjöström, T. (1996). "Interactive Implementation," Harvard Institute of Economic Research discussion paper 1751.

Chakravorty, B., Corchon, L., and Wilkie, S. (1997). "Credible Implementation," *Games Econ. Behav.* (in press).

Hurwicz, L. (1993). "Implementation and Enforcement in Institutional Modeling," in *Political Economy: Institutions, Competition and Representation* (W. Barnett, M. Hinich, and N. Schofield, Eds.). Cambridge: Cambridge University Press.

Jackson, M. (1991). "Bayesian Implementation," *Econometrica* **59**, 461–477.

Maskin, E. (1985). "The Theory of Implementation in Nash Equilibrium: A Survey," in *Social Goals and Social Organization* (L. Hurwicz, D. Schmeidler, and H. Sonnenschein, Eds.). Cambridge: Cambridge University Press.

Moore, J. (1992). "Implementation, Contracts and Renegotiation in Environments with Complete Information," in *Advances in Economic Theory: Sixth World Congress* (J.-J. Laffont, Ed.), (vol. I). Cambridge: Cambridge University Press.

Moore, J., and Repullo, R. (1990). "Nash Implementation: A Full Characterization," *Econometrica* **58**, 1083–1100.

Palfrey, T. (1992). "Implementation in Bayesian Equilibrium," in *Advances in Economic Theory: Sixth World Congress* (J.-J. Laffont, Ed.), (vol. I). Cambridge: Cambridge University Press.

Palfrey, T., and Srivastava, S. (1989). "Implementation with Incomplete Information in Exchange Economies," *Econometrica* **57**, 115–134.

Posner, R. A. (1993). "What do Judges and Justices Maximize?" Mimeo, University of Chicago Law School.

Postlewaite, A., and Schmeidler, D. (1986). "Implementation in Differential Information Economies," *J. Econ. Theory* **39**, 14–33.

Schwartz, A. (1992). "Legal Contract Theories and Incomplete Contracts," in *Contract Economics* (L. Werin and H. Wijkander, Eds.). Oxford: Basil Blackwell.

Sjöström, T. (1991). "On the Necessary and Sufficient Conditions for Nash Implementation," *Social Choice Welfare* **8**, 333–340.