JOURNAL OF ECONOMIC THEORY 77, 15–33 (1997)
ARTICLE NO. ET972318

The Theory of Implementation When the Planner Is a Player*

Sandeep Baliga

King's College, Cambridge University, Cambridge CB2 1ST, England

Luis C. Corchon

Departamento de Fundamentos, Universidad de Alicante, Alicante 03071 Spain

and

Tomas Sjöström

Department of Economics, Harvard University, Cambridge, Massachusetts 02138

Received December 13, 1995; revised April 4, 1997

In this paper we study a situation where the planner cannot commit to a mechanism and the outcome function is substituted by the planner herself. We assume (i) agents have complete information and play simultaneously and (ii) given the messages announced by the agents, the planner reacts in an optimal way given her beliefs. This transforms the implementation problem into a signaling game. We derive necessary and sufficient conditions for interactive implementation under different restrictions on the planner's out-of-equilibrium beliefs. We compare our results to standard results on Nash implementation. *Journal of Economic Literature* Classification Numbers: C72, D71, D82. © 1997 Academic Press

1. INTRODUCTION

A number of agents share some information, called the preference profile, type or state. An outside party, the principal (also called designer or planner) wants to elicit the information from the agents in order to implement an outcome that is optimal for her in each possible state (the social choice rule). In the standard approach to this problem, known as implementation,

* Many thanks to Eric Maskin, the associate editor, and an anonymous referee for their comments. We are also grateful to seminar audiences at Harvard University, Yale University, Brown University, Cambridge University, Lund University, University of Copenhagen, University of Stockholm, University of Alicante, University of Warwick, University of Windsor, and the Social Choice and Welfare meeting at Maastricht. Any remaining error are our responsibility.

0022-0531/97 $25.00

the principal can design a mechanism, i.e., a message space and an outcome function mapping messages into allocations. Once such a task has been accomplished the implementation problem becomes completely mechanical. Agents learn the state of nature, send the corresponding equilibrium message and duly receive a certain allocation. In fact, the task of the mechanism can be performed by a machine or by a mindless servant.

With the development of the theory of implementation came an appreciation of an unsatisfactory aspect on which the theory relied: out-of-equilibrium message profiles may lead to highly undesirable allocations. If the planner can irrevocably commit to the mechanism, and also prevent ex post renegotiation among agents, such bad allocations are credible. However, such assumptions are not universally regarded as satisfactory, so it is desirable to explore the consequences of assuming otherwise.

Our approach is in the spirit of Becker [4].[1] The principal is a full-fledged player who at each node of the game tree must maximize her expected payoff, so "incredible threats" of choosing very bad outcomes following certain message profiles are ruled out. Thus, in our theory of *interactive implementation* the notion of a mechanism (with its connotations of a mechanical interaction between agents and the planner) is replaced by a *cheap talk game* where in the first stage the agents simultaneously send messages, and in the second stage the planner reacts in a way that maximizes her expected utility, given her preferences and her beliefs.

With at least three agents with symmetric information, there always exists a truth telling perfect Bayesian equilibrium (PBE) of the cheap talk game. This is similar to the situation in standard Nash implementation, where "incentive compatibility" is trivially satisfied with three or more agents, and the main problem (as discussed by Maskin [9]) is to knock out undesirable equilibria. In the cheap talk game there will always exist undesirable "babbling" (or pooling) perfect Bayesian equilibria where messages do not convey (all) private information. Since we insist on full implementation, i.e., all equilibria should be optimal for the planner, some refinement is needed. We use a version of Farrell's [7] neologism proof equilibrium (see also Grossman and Perry [8] and Maskin and Tirole [11]). The corresponding notion of implementation is *interactive implementation in FGP* (*Farrell-Grossman-Perry*) *equilibrium*. The basic idea is that in a

[1] Becker [4] considered moral hazard (with observable actions) rather than adverse selection. The "rotten kid" theorem states that in a family ruled by a benevolent father who treats each family member's welfare as a "normal good," each (selfish) member is guided toward maximization of family welfare, without any need for "incredible threats." In a similar spirit, Sen [14] argued that if the head of the family is egalitarian, each member is led to equate the interests of other members with his own, making it unnecessary to precommit to an incentive scheme. In an interesting recent contribution, Ray and Ueda [13] show how the degree of egalitarianism is related to incentives to work in a team production model.

pooling equilibrium, by "objecting" (sending a zero probability message) in an credible way, an agent might be able to "convince" the planner that he is truthfully revealing some (new) information.

We find a necessary and sufficient condition for interactive implementation in FGP equilibrium. Our results can be related to the standard notion of implementation in the sense of Maskin [9]. In Maskin's model, the social optimum is given by a social choice rule. We interpret the social choice rule as representing the utility maximizing outcomes for the planner. As we show by example, even if a social choice rule is Nash implementable in the usual sense,[2] there may not exist *any* preference ordering for the planner which makes interactive implementation of the social choice rule possible. This should not be surprising, since in our model the planner cannot make incredible threats, whereas the occurrence of "incredible" outcomes out of equilibrium can be crucial in Maskin's model. On the other hand, there are social choice rules that can be interactively implemented but cannot be Nash-implemented in the standard sense. This is because when the planner is a player, her response to a given set of messages can depend on the actual equilibrium being played. (The problem of the planner's equilibrium knowledge is further studied in Baliga and Sjöström [2]).

Finally some remarks on related literature. The paper closest in spirit to ours is Chakravorty, Corchon and Wilkie [5]. They assume the person in charge of running the mechanism is a benevolent (but mindless) "keeper" and not a player: she is neither allowed to figure out the equilibrium strategies of the agents nor to make inferences from the messages sent by the agents. Therefore, she is always uninformed. But she must keep in the spirit of the mechanism and therefore under no circumstances can she pick an allocation that is not in the range of the social choice rule. In a slightly different attack on the credibility problem, Maskin and Moore [10] assumed that *the agents* cannot commit not to renegotiate the outcome recommended by the mechanism. Their approach is relevant if no principal is present, and the mechanism is a sort of constitution for the agents.

2. EXAMPLES

We first give two examples illustrating why interactive implementation in FGP equilibrium is in general neither easier nor more difficult than standard Nash implementation. For examples 1 and 2, we suppose there are three consumers and three commodities, and two states of the world θ'

[2] In the exchange economy, this is equivalent to the well-known condition of Maskin monotonicity.

and θ''. However, the third consumer is only interested in the consumption of the third good and no other consumer has endowments of this commodity nor do they derive any utility from consuming the third good. We assume the third consumer always consumes just her initial endowments so we in effect have a two-good, two consumer world.

EXAMPLE 1. A Social Choice Rule which is not interactively implementable in FGP-equilibria, even though it is Nash-implementable.

Let ϕ be the Walrasian correspondence. In Fig. 1 we show the competitive equilibria for the states θ' and θ'' together with agent 1's indifference curves. The outcomes $\phi(\theta')$ and $\phi(\theta'')$ are the most preferred outcomes by the planner in states θ' and θ'', respectively. Consider the cheap talk game where all three agents simultaneously send messages to the principal. The message space is sufficiently big to at least include all subsets of the states of the world. After the agents have spoken, the planner picks an allocation. There exists a truth-telling separating perfect Bayesian equilibrium, where the planner's off the equilibrium path beliefs are such that if one agent should deviate, the planner believes the majority tells the truth. Any separating equilibrium reveals the true state to the planner and allows her to pick the right outcome in each state. Now consider a non-revealing (pooling) perfect Bayesian equilibrium where the agents say the same thing in each state (say the agents always claim that the state is θ''). For any message, the planner's prior beliefs go through and she picks $\hat{a}$, the *optimal compromise*: if she cannot get any information from the agents then she

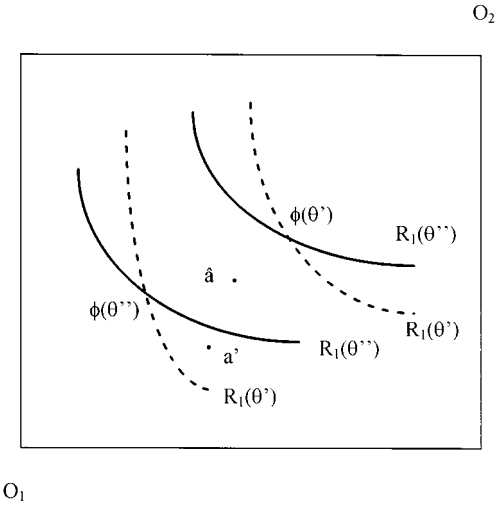

FIG. 1. Example 1.

prefers $\hat{a}$. (It is clear that a utility function for the planner rationalizing this choice exists). An *objection* is a zero probability message under the equilibrium strategies.[3] Suppose the state is truly θ' and agent 1 objects: "Please implement $\phi(\theta')$ and not $\hat{a}$ as the state is truly θ'." This objection is not reliable: agent 1 prefers $\phi(\theta')$ to $\hat{a}$ in both states, so this speech should not convince the planner that the state is θ' (Farrell [7]). As agent 2 certainly has no incentive to convince the planner that the state is θ', and as the situation in state θ'' is symmetric, the pooling equilibrium is an FGP equilibrium, i.e., an equilibrium which is free from reliable objections. Therefore, the competitive equilibrium cannot be interactively implemented.[4]

On the other hand, if the planner could commit, then there exists a "canonical" mechanism which can Nash implement this (Maskin-monotonic) ϕ (see Osborne and Rubinstein [12, Section 10.4]). If all agents announce θ'' always, the canonical mechanism would always pick $\phi(\theta'')$, but agent 1 can "object" by asking for an allocation such as a' (see Fig. 1), which he prefers to $\phi(\theta'')$ if the state is θ' but not if the state is θ''. The usual interpretation is that by making this objection, agent 1 "persuades the planner that the preference relation announced for him by the others is incorrect" (Osborne and Rubinstein [12, p. 188]) The condition of Maskin-monotonicity implies that such "persuasive objections" exist. In our model, the situation is clearly different. If the agents always announce θ'', the planner "knows it" and will respond with $\hat{a}$ rather than $\phi(\theta'')$, and moreover an outcome such as a' is totally irrelevant unless it is a best response for the planner against some beliefs.

EXAMPLE 2. An SCR ϕ which is interactively implementable in FGP equilibrium, even though it is not Nash implementable.

Figure 2 shows the Walrasian correspondence for states θ' and θ'' together with agent 1's indifference curves. In this case this correspondence does not satisfy Maskin-monotonicity and thus it is not Nash implementable in the standard sense. Again, $\hat{a}$ denotes the "optimal compromise." Consider a pooling equilibrium where no information is revealed. In state θ'' agent 1 can object that the state is truly θ'', and this is a convincing speech because in state θ'' he prefers $\phi(\theta'')$ to $\hat{a}$, while in state θ' he prefers $\hat{a}$ to $\phi(\theta'')$. This

[3] To make sure that objections are always available, we include auxiliary messages such as "integers." Still, if the agents would "babble" by sending each message with positive probability, no zero-probability message might exist. However, we suppose such mixed strategies are not used.

[4] More precisely, we have shown that these message spaces do not work. It is clear that no other message space will work either.

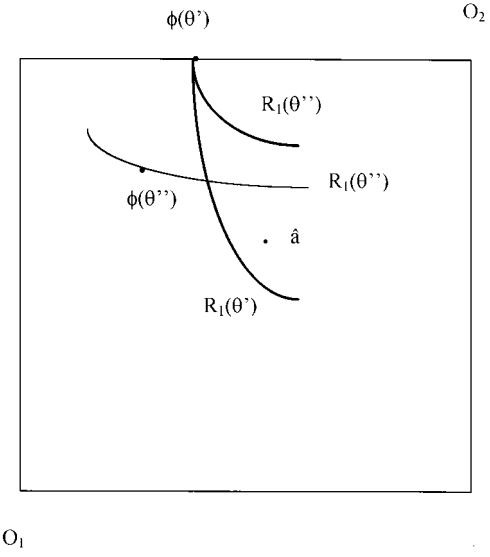

FIG. 2. Example 2.

breaks the pooling equilibrium. As a separating equilibrium always exists and is optimal, ϕ is interactively implementable.[5]

Subtle issues arise in our model when there are more than two possible states. This is illustrated by our next example.

EXAMPLE 3. There are three agents $I = \{1, 2, 3\}$ and three states $\Theta = \{\alpha, \beta, \gamma\}$. The agents' message spaces include at least all possible subsets of states. The ranking of the outcomes in the different states by agent 1 and the planner are as follows:

	Agent 1			Planner		
State:	α	β	γ	α	β	γ
	a	*b*	*b*	*a*	*b*	*c*
	d	*d*	*a*	*d*	*d*	*d*
	c	*c*	*c*	*c*	*c*	*b*
	b	*a*	*d*	*b*	*a*	*a*

[5] The "canonical" mechanism for standard Nash implementation fails because it can get "stuck" at $\phi(\theta')$: any outcome which an agent prefers to $\phi(\theta')$ when the state is θ'' would also be preferred when the state is θ'.

The three states are equally likely. The planner's utility function is such that relative to beliefs that put probability 1/2 each on α and β and zero on γ, her best response is d.

Consider a perfect Bayesian equilibrium where all three agents announce $\{\alpha, \beta\}$ in states α and β and $\{\gamma\}$ in state γ. The planner picks c if at least two agents say $\{\gamma\}$, d if at least two agents say $\{\alpha, \beta\}$, and otherwise plays a best response to some arbitrary beliefs. But agent 1 can make a reliable "speech" in state β: "you know from *the other agents' messages* $\{\alpha, \beta\}$ that the state is definitely not γ. It is really β, so choose b. I have no reason to argue this if it is α, but if it is β I do have this incentive, so you should believe me." So the equilibrium is not FGP. Now consider the following perfect Bayesian equilibrium. Agents 2 and 3 always announce Θ; agent 1 announces $\{\alpha, \beta\}$ in states α and β and $\{\gamma\}$ in state γ. The planner picks c if agent 1 says $\{\gamma\}$ and at least one other agent says Θ, and d if she hears any other message profile (the latter is supported by the belief that the state is α or β with equal probability). As in the equilibrium above, in both states α and β the outcome is d, and agent 1 prefers b (the planner's best response to β) to d in state β, but d to b in state α. But now agent 1 cannot convince the principal that the state is β, because the other two agents' messages do not allow the planner to rule out state γ, and in state γ outcome b is agent 1's favorite. Can agent 1 in state β convince the principal that the state is in the set $\{\beta, \gamma\}$? This clearly depends on *what the planner would do* if she became convinced that the state is in the set $\{\beta, \gamma\}$. In turn, this depends on what *relative probabilities* she puts on β and γ. As this is an out-of-equilibrium situation, it is not clear that the relative probabilities on $\{\beta, \gamma\}$ should be determined by the prior.

This example illustrates two issues: (1) If an objection convinces the planner that the state is in some set T, how are the relative probabilities over T determined?[6] (2) If a player makes an objection, exactly how much information can the planner obtain from *the other players'* (equilibrium) messages?

3. A GENERAL FORMULATION

There are $n \geqslant 3$ agents. Let I be the set of agents. The set of feasible outcomes is denoted by A. Let Θ be the finite set of possible states of the world. Let 2^{Θ} be the set of all subsets of Θ. The prior probability of state θ occurring is $p(\theta) > 0$ for all $\theta \in \Theta$. If $T \subseteq \Theta$, then $p(T) \equiv \Sigma_{\theta \in T} p(\theta)$. The probability distribution p_T is derived from the prior as follows: $p_T(\theta) = 0$

[6] We are grateful to an anonymous referee for stressing this Issue.

if $\theta \notin T$, and $p_T(\theta) = p(\theta)/p(T)$ if $\theta \in T$. For any set X, let $\#X$ denote the number of elements in X.

Weak preferences of agent i in state θ are given by the ordering $R_i(\theta)$. Thus, for $a, b \in A$, $aR_i(\theta)\, b$ means agent i (weakly) prefers outcome a to outcome b in state θ. Let $P_i(\theta)$ represent strict preferences and $I_i(\theta)$ indifference. The lower contour set for agent i at allocation a and state θ is $L_i(a, \theta) = \{b \in A : aR_i(\theta)\, b\}$. We assume throughout that the true state θ is common knowledge among the agents.

At this point, the literature on mechanism design defines a concept of social welfare, a social choice rule (SCR), $F: \Theta \to A$. We recall the following definition. If for all $b \in A$, $aR_i(\theta)\, b$ implies $aR_i(\theta')\, b$, then $R_i(\theta')$ is a monotonic transformation of $R_i(\theta)$ at a. The social choice rule F is (Maskin) monotonic if, whenever $a \in F(\theta)$ and for all i, $R_i(\theta')$ is a monotonic transformation of $R_i(\theta)$ at a, then $a \in F(\theta')$. Also, we say that f is a *selection* from F, and write $f \in F$, if f is a single-valued function such that $f(\theta) \in F(\theta)$ for all $\theta \in \Theta$.

In our setting, the planner is just another player, with an objective function and a strategy space. The outcomes at the top of the planner's objective function in each state can be thought of as defining the social choice rule. The planner differs from the other players in one fundamental respect: the state is common knowledge to them but not to her. Thus, if the agents are not using strategies that release their private information in all states, she will have to choose a best response even though she is not sure of the state.

If allocation a is chosen in state θ, the payoff to the planner is $U(a, \theta)$. The implied social choice rule is

$$F(\theta) \equiv \operatorname*{argmax}_{a \in A} U(a, \theta) \tag{1}$$

Conversely, if F is a given social choice rule and U is such that (1) holds for all θ, then U is *compatible with F*.

Let r be a probability distribution over Θ and given some $T \subseteq \Theta$ let $\Delta(T)$ be the set of probability distributions over T. Then, given the planner's utility function U, we define

$$BR(r) \equiv \operatorname*{argmax}_{a \in A} \sum_{\theta \in \Theta} r(\theta)\, U(a, \theta)$$

and for any $T \subseteq \Theta$,

$$BR(T) \equiv \bigcup_{r \in \Delta(T)} BR(r).$$

Also, for any $T \subseteq \Theta$ define $B(T) \equiv BR(p_T)$, where p_T is derived from the prior as above. Note that $BR(T)$ is the set of the principal's best responses

for *some* belief concentrated on the set T, whereas $B(T)$ are the best responses if she is not only convinced that the state is in T, but in addition assigns relative probabilities to states in T according to the prior. Finally, we say b is a *compromise selection* from B if b is a single-valued function such that $b(T) \in B(T)$ for all $T \subseteq \Theta$.

4. INTERACTIVE IMPLEMENTATION

The *message spaces* are $M = \times_{i \in I} M_i$, where $M_i = 2^{\Theta} \times Q_i$, and $Q_i = \{1, 2, ..., \#\Theta + 1\}$.[7] Thus, each agent reports a subset of states $T_i \subseteq \Theta$ and a "nuisance message" $q_i \in Q_i$. A generic message is denoted $m_i = (T_i, q_i)$. Let $m_{-i} = (m_1, ..., m_{i-1}, m_{i+1}, ..., m_n)$. A strategy for agent i is a map μ_i: $\Theta \to M_i$, where $\mu_i(\theta)$ is the message sent in state θ. Let $\mu_{-i}(\theta) = (\mu_1(\theta), ..., \mu_{i-1}(\theta), \mu_{i+1}(\theta), ..., \mu_n(\theta))$. A strategy for the planner is a function α: $M \to A$, where $\alpha(m)$ is the allocation chosen in response to the message m.

Suppose the agents use strategies μ. The range of μ is denoted $\mu(\Theta) = \{m \in M: m = \mu(\theta) \text{ for some } \theta \in \Theta\}$. For any $m \in M$, $\mu^{-1}(m) \equiv \{\theta \in \Theta: \mu(\theta) = m\}$ is the set of states where agents send message m. Similarly, $\mu_i^{-1}(m_i) \equiv \{\theta \in \Theta: \mu_i(\theta) = m_i\}$ and $\mu_{-i}^{-1}(m_{-i}) \equiv \{\theta \in \Theta: \mu_{-i}(\theta) = m_{-i}\}$. If $\mu(\theta) = m$ for all $\theta \in T \subseteq \Theta$, we write $m = \mu(T)$. Similarly if $\mu_i(\theta) = m_i$ for all $\theta \in T \subseteq \Theta$, then $m_i = \mu_i(T)$, and if $\mu_{-i}(\theta) = m_{-i}$ for all $\theta \in T \subseteq \Theta$, then $m_{-i} = \mu_{-i}(T)$.

DEFINITION 1. (μ^*, α^*) is a *perfect Bayesian equilibrium* (PBE) if

(1) for each $\theta \in \Theta$ and each i, $\alpha^*(\mu^*(\theta))\, R_i(\theta)\, \alpha^*(\mu^*_{-i}(\theta), m_i)$ for all $m_i \in M_i$,

(2) for each $m \in \mu^*(\Theta)$, $\alpha^*(m) \in BR(p_T)$, where $T = (\mu^*)^{-1}(m)$,

(3) for each $m \in M \setminus \mu^*(\Theta)$, there exists $r \in \Delta(\Theta)$ such that $\alpha^*(m) \in BR(r)$.

Part (1) of Definition 1 states that, given the anticipated response from the planner, each agent sends a message that maximizes his payoff. Part (2) requires that, for each equilibrium message m, the planner chooses what is best for her, conditional on the *correct* belief that the true state belongs to $(\mu^*)^{-1}(m)$. Part (3) requires that if m is *not* sent in equilibrium, then there exists *some* belief for the planner such that the planner's response is optimal conditional on this belief.

[7] We can consider more general message spaces but nothing is lost by focusing our attention on the ones we consider.

A PBE is *separating* if $(\mu^*)^{-1}(m)$ is a singleton for all $m \in \mu^*(\Theta)$. In this case the planner can invert μ^* and is fully informed in equilibrium. A PBE which is not separating is *pooling*. In a pooling PBE some "compromise" must be chosen by the planner whenever $m \in \mu^*(\Theta)$ is such that $(\mu^*)^{-1}(m)$ is not a singleton.

Even though all agents have the same information, in equilibrium a *single agent* may reveal some unique information (as happened in Example 3). Such equilibria must be "incentive-compatible" (in the second equilibrium of Example 3, agent 1 prefers d to c in states α and β, but c to d in state γ). It will be useful to formalize this notion. A *deception for agent* i, δ_i, is a mapping from Θ to 2^Θ which satisfies $\theta \in \delta_i(\theta)$ for all θ. A *deception*, denoted δ, is a profile of deceptions, one for each agent. Let $\delta(\theta) = \bigcap_{i \in I} \delta_i(\theta)$, $\delta_{-j}(\theta) = \bigcap_{i \neq j} \delta_i(\theta)$.

If the agents use strategies μ, then this implies a deception defined by, for each i,

$$\delta_i(\theta) \equiv \mu_i^{-1}(\mu_i(\theta)) \tag{2}$$

Then $\delta(\theta)$ are the states the principal, knowing μ, will not be able to distinguish from θ. Also, (2) implies $\delta_{-i}(\theta) = \mu_{-i}^{-1}(\mu_{-i}(\theta))$, i.e., $\delta_{-i}(\theta)$ are the states the principal cannot distinguish from θ by looking at the messages sent by all agents except i. If $\theta' \in \delta_{-i}(\theta)$ and $\delta(\theta) \neq \delta(\theta')$ (where the δ_i are still defined as in (2)), then δ is a finer partition of the states than δ_{-i}, that is, under strategy profile μ agent i is revealing some unique information. If $T_i \subseteq \Theta$ for all i, define $D(T_1, ..., T_n) \equiv \{j \in I: \bigcap_{i \in I} T_i = \varnothing, \bigcap_{i \neq j} T_i \neq \varnothing\}$. If (for each i) $T_i = \mu_i^{-1}(m_i)$ for some equilibrium message m_i, then $i \in D(T_1, ..., T_n)$ means the messages of the agents in $I \setminus \{i\}$ are mutually consistent, given the equilibrium strategies, but the whole n-tuple of messages is inconsistent. If $\# D(T_1, ..., T_n) > 1$, then it is not possible to single out a unique agent as being inconsistent with the others.[8]

DEFINITION 2. Given a compromise selection b, a deception δ is *incentive compatible with respect to* b if the following holds:

(i) for all $i \in I$ and $\theta \in \Theta$

$$b(\delta(\theta))\, R_i(\theta)\, b(\delta(\theta')) \qquad \text{for all} \quad \theta' \in \delta_{-i}(\theta).$$

(ii) For all $(\theta_1, \theta_2, ..., \theta_n) \in \Theta^n$, if $\# D(T_1, ..., T_n) > 1$ where for each i, $T_i = \delta_i(\theta_i)$, then there exists $p' \in \Delta(\Theta)$ and $a = a(T_1, ..., T_n) \in BR(p')$ such that for each $i \in D(T_1, ..., T_n)$ and each $\theta \in \bigcap_{j \neq i} T_j$, $b(\delta(\theta))\, R_i(\theta)\, a$.

[8] In particular, if in equilibrium only two agents reveal information, but one of them were to deviate so their reports contradict each other, the principal may not be able to figure out who has deviated.

If $\delta_i(\theta) \equiv \mu_i^{-1}(\mu_i(\theta))$ is a deception corresponding to some equilibrium (where the principal breaks ties according to a compromise selection b), then (as the proof of Theorem 2 below shows), δ must be incentive compatible. Indeed, part (i) of Definition 2 implies maximization on behalf of each agent i who reveals unique information on his own. Part (ii) is a condition similar to the conditions that guarantee incentive compatibility in standard 2-person Nash implementation.

Now we introduce restrictions on the planner's off-the-equilibrium-path beliefs in the spirit of Farrell [7]. However, in contrast to the standard models, there is more than one "sender." Therefore, if one agent makes a surprise announcement, the planner may infer some information from the other agents' messages.

DEFINITION 3. Let (μ^*, α^*) be a PBE. Suppose there exists θ' such that $\mu^*_{-i}(\theta') = m_{-i}$, $(T', q') \in M_i$, $T' \subset (\mu^*)^{-1}_{-i}(m_{-i})$ but $(m_{-i}, (T', q')) \notin \mu^*(\Theta)$. Then (T', q') is an *objection* to m_{-i} by player i.

Thus, in some particular equilibrium the agents send $m = \mu^*(\theta')$ in state θ'. Observing m_{-i}, the principal infers that the state is in $(\mu^*)^{-1}_{-i}(m_{-i})$. An objection from agent i is a deviation from his equilibrium strategy signaling that the state is truly in the set $T' \subset (\mu^*)^{-1}_{-i}(m_{-i})$.

DEFINITION 4. Let (μ^*, α^*) be a PBE, and $\mu^*_{-i}(\theta') = m_{-i}$. An objection (T', q') to m_{-i} is *BR-reliable* for player i if, $T' \subset (\mu^*)^{-1}_{-i}(m_{-i})$, and

(1) for all $\theta \in T'$ and all $a' \in BR(T')$, $a' P_i(\theta)\, \alpha^*(\mu_i^*(\theta), m_{-i})$, and

(2) for all $\theta \in (\mu^*)^{-1}_{-i}(m_{-i}) \backslash T'$ and all $a' \in BR(T')$, $\alpha^*(\mu_i^*(\theta), m_{-i})\, R_i(\theta)\, a'$.

A BR-reliable objection amounts to the following speech: "The other agents have announced $m_{-i} = \mu^*_{-i}(\theta')$ but I object to that: the state is truly in T' and you should pick some element a' in $BR(T')$. Your knowledge of strategies and the other agents' messages tells you that the true state is in $(\mu^*)^{-1}_{-i}(m_{-i})$. But now notice that the set $T' \subset (\mu^*)^{-1}_{-i}(m_{-i})$ satisfies (1) and (2) of Definition 4. Given this, I will have the incentive to object iff $\theta \in T'$, so my speech is credible."

Definition 4 supposes that, once the principal is convinced that the state is in the set T', any probability distribution with support T' might be a candidate for the ex post beliefs, consistency requires that the set T' is precisely the set of types that would profit for any belief concentrated on T'. On the other hand, Farrell [7] assumes that if a message convinces the planner that the state is in T', her ex post beliefs are given by the priors restricted to the set T', denoted $P_{T'}$ (see also Maskin and Tirole [11]).

Accordingly, in Definition 4 we can replace the set $BR(T')$ by the set $B(T') \equiv BR(p_{T'})$ to get the definition of B-reliable objection.[9]

DEFINITION 5. A PBE is a *weak FGP-equilibrium* if no player has a BR-reliable objection against any message which is sent in equilibrium with positive probability. A PBE is an *FGP-equilibrium* if no player has a B-reliable objection against any message which is sent in equilibrium with positive probability.

DEFINITION 6. The social choice rule F (as defined by (1)) is (interactively) *implemented in weak FGP-equilibrium* (resp. FGP-equilibrium) if:

(i) for each selection $f \in F$, there exists a weak FGP equilibrium (resp. FGP equilibrium) (μ, α) such that $\alpha(\mu(\theta)) = f(\theta)$ for all θ; and

(ii) if (μ, α) is a weak FGP-equilibrium (resp. FGP-equilibrium), then for all θ, $\alpha(\mu(\theta)) \in F(\theta)$.

Because there always exist truth-telling equilibria and these are trivially FGP equilibria, only part (ii) of Definition 6 has bite. And since any BR-reliable objection is B-reliable, it is easier to knock out pooling equilibria using B-reliable objections (hence an FGP equilibrium is also a weak FGP equilibrium). Thus, any SCR which is implemented in weak FGP equilibria is also interactively implementable in FGP equilibria. When there are only two states, the two concepts are equivalent (in this case an objection can only be made by singletons, and if $T = \{\theta\}$ then $B(T) = BR(T)$).

Similarly to Definition 6, one can define interactive implementation in PBE (with no restrictions on out-of-equilibrium beliefs). But due to the existence of "babbling" PBE, interactive implementation in PBE is (almost) impossible.

THEOREM 1. *If F is interactively implementable in PBE, then there exists an outcome a such that $a \in F(\theta)$, $\forall \theta \in \Theta$.*

Proof. Suppose F is interactively implemented in PBE using message spaces $\times_{i \in I} M_i$. Let all agents send the message profile m independent of the state of the world. For any message, the planner's prior goes through and she implements some $a \in BR(p)$, where p is the prior belief. These strategies and posteriors form a pooling PBE, and since F is implemented, we must conclude that $a \in F(\theta)$, $\forall \theta \in \Theta$. Q.E.D.

[9] This method of proceeding shows that our general approach can be used together with many different assumptions about the principal's out-of-equilibrium beliefs.

5. NECESSARY AND SUFFICIENT CONDITIONS FOR INTERACTIVE IMPLEMENTATION

We introduce a condition which guarantees that reliable objections can be used to knock out any pooling equilibrium. In our setting, this condition replaces Maskin monotonicity. As is to be expected, the condition depends on the planner's utility function U, but only through the sets $B(T)$ and $BR(T)$.

DEFINITION 7. The social choice rule F, defined by Eq. (1), is *weakly reliably monotonic* if the following holds. Suppose that a deception δ is incentive compatible with respect to some compromise selection b and there exists a state θ such that $b(\delta(\theta)) \notin \bigcap_{t \in \delta(\theta)} F(t)$. Then, there exists $i \in I$, $\theta' \in \Theta$, $T' \subset S \equiv \delta_{-i}(\theta')$ such that:

(i) if $\theta \in T'$ then $aP_i(\theta)\, b(\delta(\theta))$ for all $a \in BR(T')$

(ii) if $\theta \in S \backslash T'$, then $b(\delta(\theta))\, R_i(\theta) a$ for all $a \in BR(T')$

THEOREM 2. *F is implementable in weak FGP equilibrium if and only if F is weakly reliably monotonic.*

Proof. Necessity: Suppose F is interactively implementable in weak FGP equilibrium using message spaces $\times_{i \in I} M_i$, where for each i, $M_i = 2^{\Theta} \times Q_i$. Suppose for some compromise selection b the deception δ is incentive compatible with respect to b and there is some state θ such that $b(\delta(\theta)) \notin \bigcap_{t \in \delta(\theta)} F(t)$.

Consider the following perfect Bayesian equilibrium (μ^*, α^*). For all $i \in I$ and $\theta \in \Theta$, $\mu_i^*(\theta) = (\delta_i(\theta), q_i^*)$ where q_i^* does not depend on θ. By construction, if $m = \mu^*(\theta)$ then $(\mu^*)^{-1}(m) = \delta(\theta)$ and $(\mu^*)_{-i}^{-1}(m_{-i}) = \delta_{-i}(\theta)$ for all i. For message profile m, with $m_i = (T_i, q_i)$, define the principal's response $\alpha^*(m)$ as follows. (i) If $m \in \mu^*(\Theta)$, then $\alpha^*(m) = b((\mu^*)^{-1}(m))$. (ii) If there is θ and i such that $m_j = \mu_j^*(\theta)$ for all $j \neq i$, and either $m_i \notin \mu_i^*(\Theta)$ or $D(T_1, ..., T_n) = \{i\}$ then set $\alpha^*(m) = b(\delta(\theta'))$ for some $\theta' \in \delta_{-i}(\theta)$. (iii) If for each i there is θ_i such that $m_i = \mu_i^*(\theta_i)$ and $\# D(\delta_1(\theta_1), ..., \delta_n(\theta_n)) > 1$, then pick $\alpha^*(m) = a(\delta_1(\theta_1), ..., \delta_n(\theta_n))$ as defined by Definition 2 part (ii). (iv) For all other m, $\alpha^*(m) \in BR(p')$ for some arbitrary $p' \in \Delta(\Theta)$.

According to (i) the planner is maximizing his utility relative to equilibrium messages, and according to (ii)–(iv) there exists beliefs that support his reaction to out-of-equilibrium messages. As the deception δ is incentive compatible with respect to b, the agents are also optimizing. For, suppose at state θ player i deviates to $m_i \neq \mu_i^*(\theta)$. Each other player $j \neq i$ is sending $\mu_j^*(\theta) = (\delta_j(\theta), q_j^*) = (T_j, q_j^*)$. If $m_i = \mu_i^*(\theta')$ for some $\theta' \in \delta_{-i}(\theta)$ then

player i is not better off by Definition 2 part (i). The same is true if either $m_i \notin \mu_i^*(\Theta)$, or $m_i = (T_i', q_i^*) \in \mu_i^*(\Theta)$ with $D(T_{-i}, T_i') = \{i\}$ (where $(T_{-i}, T_i') \equiv (T_1, ..., T_{i-1}, T_i', T_{i+1}, ..., T_n)$) For in this case $\alpha^*(m) = b(\delta(\theta'))$ for some $\theta' \in \delta_{-i}(\theta)$, which is an outcome agent i could have attained by sending $\mu_i^*(\theta')$. Finally, suppose $m_i = (T_i', q_i^*) = (\delta_i(\theta'), q_i^*) = \mu_i^*(\theta')$ for some $\theta' \notin \delta_{-i}(\theta)$ and $\# D(T_{-i}, T_i') > 1$. Then $i \in D(T_{-i}, T_i')$ and $b(\delta(\theta))\ R_i(\theta)\ \alpha^*(m)$ by Definition 2 part (ii). Thus, (μ^*, α^*) is a perfect Bayesian equilibrium.

Since F is implemented and there exists a state θ such that $\alpha^*(\mu^*(\theta)) = b(\delta(\theta)) \notin \bigcap_{t \in \delta(\theta)} F(t)$, (μ^*, α^*) is not a weak FGP equilibrium. Therefore, some agent i in some state θ' must have a BR-reliable objection (T', q) to $\mu_{-i}^*(\theta')$. It must be the case that $T' \subset (\mu^*)_{-i}^{-1}(\mu_{-i}^*(\theta')) = \delta_{-i}(\theta') \equiv S$, and the following holds: if $\theta \in T'$, $aP_i(\theta)\ \alpha^*(\mu^*(\theta))) = b(\delta(\theta))$ for all $a \in BR(T')$; if $\theta \in S \backslash T'$, $\alpha^*(\mu^*(\theta))) = b(\delta(\theta))\ R_i(\theta)\ a$ for all $a \in BR(T')$. Thus, F is weakly reliably monotonic. This proves necessity.

Sufficiency: Let the message space for player i be $M_i = 2^{\Theta} \times \{1, 2, ..., |\Theta| + 1\}$. Truthtelling can be supported as an FGP-equilibrium by letting the planner disregard unilateral deviations. Thus, we only need to show that there are no non-optimal equilibria.

Suppose there exists a non-optimal weak FGP equilibrium (μ, α) such that for some $\theta^* \in \Theta$, $\alpha(\mu(\theta^*)) \notin F(\theta^*)$. Define a deception δ as follows: for all i in I and all θ, $\delta_i(\theta) = \mu_i^{-1}(\mu_i(\theta))$. Notice that $\delta(\theta) = \mu^{-1}(\mu(\theta))$ for all θ. Define a compromise selection b as follows: for all $T \subseteq \Theta$ such that $T = \delta(\theta)$ for some $\theta \in \Theta$, set $b(T) = \alpha(\mu(\theta))$. Otherwise, $b(T)$ is arbitrary.

We claim δ is incentive compatible with respect to b. For part (i) of Definition 2, notice that for all $\theta' \in \mu_{-i}^{-1}(\mu_{-i}(\theta)) = \delta_{-i}(\theta)$, as (μ, α) is a perfect Bayesian equilibrium,

$$b(\delta(\theta)) = \alpha(\mu(\theta))\ R_i(\theta)\ \alpha(\mu_i(\theta')), \mu_{-i}(\theta)) = \alpha(\mu(\theta')) = b(\delta(\theta'))$$

For part (ii) of Definition 2, suppose $(T_1, ..., T_n) = (\delta_1(\theta_1), ..., \delta_n(\theta_n))$ satisfies $\# D(T_1, ..., T_n) > 1$. Let $m \equiv (\mu_1(\theta_1), ..., \mu_n(\theta_n))$. If $i \in D(T_1, ..., T_n)$, then $\bigcap_{j \neq i} T_j \neq \varnothing$. Now, if $\theta \in T_j$ then $\mu_j(\theta) = \mu_j(\theta_j)$ by definition. Therefore, if $\theta \in \bigcap_{j \neq i} T_j$ we have $m_{-i} = \mu_{-i}(\theta)$ and $m = (\mu_{-i}(\theta), m_i)$. As (μ, α) is a perfect Bayesian equilibrium,

$$\alpha(\mu(\theta)) = b(\delta(\theta))\ R_i(\theta)\ \alpha(\mu_{-i}(\theta), m_i) = \alpha(m)$$

Now set $a(T_1, ..., T_n) = \alpha(m)$ to get part (ii) of Definition 2. Hence, the deception δ is incentive compatible with respect to b.

Since F is weakly reliably monotonic, there exists $i \in I$, $\theta' \in \Theta$, $T' \subset \delta_{-i}(\theta') = S$ such that: if $\theta \in T'$ then $aP_i(\theta)\ b(\delta(\theta))$ for all $a \in BR(T')$, and if $\theta \in S \backslash T'$, then $b(\delta(\theta))\ R_i(\theta)\ a$ for all $a \in BR(T')$. Let the state be $\theta \in T'$ and consider $m_i' = (T', z) \notin \mu^*(\Theta)$. Then, m_i' is a BR-reliable objection to

$m_{-i} = \mu_{-i}(\theta)$. Then (μ, α) is not a weak FGP equilibrium, a contradiction. This proves sufficiency. Q.E.D.

By replacing "$a \in BR(T')$" by "$a \in B(T')$" in Definition 7, we get the corresponding definition of *reliably monotonic* and the following result. We omit the proof as it similar to that of Theorem 2 (see Baliga, Corchon and Sjöström [1]).

THEOREM 3. *The social choice rule F is implementable in FGP equilibrium if and only if it is reliably monotonic.*

The examples of Section 2 show that Maskin monotonicity is neither necessary nor sufficient for interactive implementation in FGP equilibria. They also highlight the role played by "compromise" alternatives. Given a Maskin monotonic social choice function F, we may wonder if there *always* exists *some* preferences for the planner which are compatible with F and allow interactive implementable in FGP equilibria. The answer is no. We can exhibit a Maskin monotonic social choice function F such that, if the planner's utility function is *any* utility function compatible with F, F cannot be interactively implemented it FGP equilibria (and a fortiori not in weak FGP equilibria).

EXAMPLE 4. A Maskin-monotonic social choice rule F such that, if the planner's utility function U is any utility function which is compatible with F, F cannot be interactively implemented in FGP equilibria.

Consider a three person exchange economy with two goods. The social endowment of good i is ω_i. There are four states, $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. The preferences of player 3 are fixed at $R_3(\theta) = R_3$ for all θ. The preferences of player 1 are $R_1(\theta_1) = R_1(\theta_2) = R_1$ and $R_1(\theta_3) = R_1(\theta_4) = R_1'$. The preferences of player 2 are $R_2(\theta_1) = R_2(\theta_3) = R_2$ and $R_2(\theta_2) = R_2(\theta_4) = R_2'$. Let $a = F(\theta_1)$, $b = F(\theta_2)$, $c = F(\theta_3)$, $d = F(\theta_4)$ be four distinct outcomes. Suppose in all four cases player 3 gets some small amount $\epsilon > 0$ of each good. Let $x_i = (x_{i1}, x_{i2})$ denote the amount of goods 1 and 2 consumed by agent i at allocation x. The preferences of player 1 are given in Fig. 3, where the dotted (resp. solid) line represents an R_1' (resp. R_1) indifference curve. The preferences of player 2 are given in Fig. 4. R_1' and R_2' are actually isomorphic, and also R_1 and R_2. The indifference curves are drawn such that both player 1 and player 2 are always indifferent between a, b, c and d. (But the example can be perturbed so that this indifference goes away.)

As F is Maskin monotonic, it is Nash implementable in the standard sense. Suppose the planner's preferences are represented by U, where U is any utility function compatible with F. We claim F cannot be reliably monotonic.

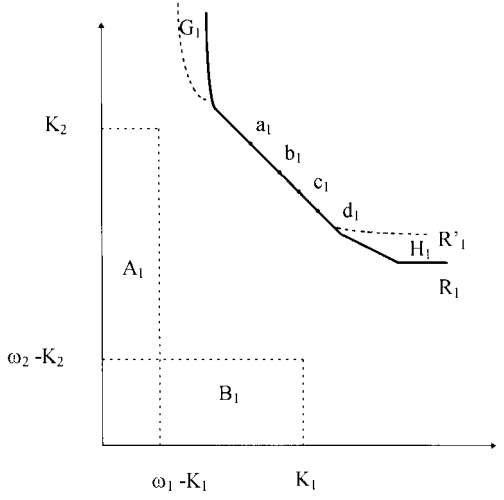

FIG. 3. Player 1's preferences.

Let $d_{11} = a_{21}$ be the greatest amount of good 1 consumed by any player at any of the outcomes a, b, c, d, and let $a_{12} = d_{22}$ denote the greatest amount of good 2 consumed by any player at any of the outcomes a, b, c, d. Let K_1 and K_2 be numbers such that $d_{11} < K_1 < \omega_1$ and $a_{12} < K_2 < \omega_2$. Let the area $C_1 \equiv A_1 \cup B_1$ in Fig. 3 be the union of the sets $A_1 = \{x_1 : x_{11} \leqslant \omega_1 - K_1$ and $x_{12} \leqslant K_2\}$ and $B_1 = \{x_1 : x_{11} \leqslant K_1$ and $x_{12} \leqslant \omega_2 - K_2\}$. Let $C_2 \equiv A_2 \cup B_2$ in Fig. 4 be similarly defined.

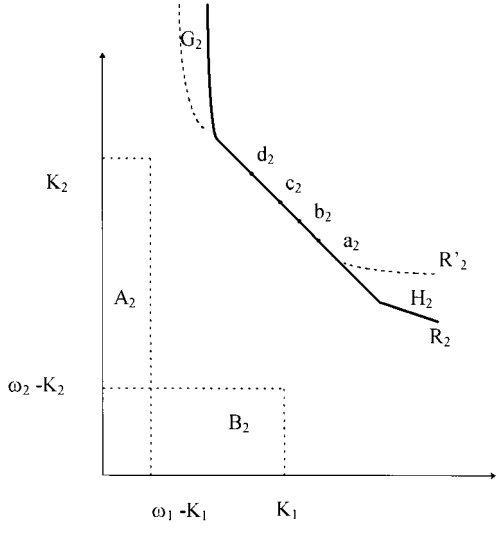

FIG. 4. Player 2's preferences.

Now we draw the indifference curves for player 1 in such a way that if an indifference curve for preferences R_1 passes through the area C_1, then it coincides throughout the consumption set with an indifference curve for preferences R_1'. Similarly, if an indifference curves for preferences R_2 passes through the area C_2, then it coincides throughout player 2's consumption set with an indifference curve for preferences R_2'.

Let G_1 and H_1 be the areas in Fig. 3 given by:

$G_1 = \{z$: if $x \in A$ and $x_1 = z$, then $aP_1 x$ and $xR_1' a\}$

$H_1 = \{z$: if $x \in A$ and $x_1 = z$, then $aP_1' x$ and $xR_1 a\}$

Let G_2 and H_2 be similar for player 2.

It is clear that we can draw the indifference curves in such a way that if $z = (z_1, z_2) \in G_1$ (where z_i is the consumption of good i) then $z_1 > \omega_1 - K_1$ and $z_2 > K_2$. Similarly, if $z = (z_1, z_2) \in G_2$ then $z_1 > \omega_1 - K_1$ and $z_2 > K_2$. Similar statements hold for H_1 and H_2.

Suppose F is reliably monotonic and let $e \in B(\Theta)$. Consider the deception $\delta_i(\theta) = \Theta$ for all θ for all i, and also the compromise selection $b(\Theta) = e$ and $b(T)$ is arbitrary for all other $T \subseteq \Theta$. Clearly, δ is incentive compatible with respect to b. Since F is reliably monotonic and $e \notin \bigcap_{t \in \Theta} F(t)$, there exists $i \in I$, $T' \subset \Theta$ and $g \in B(T')$ such that:

(i) if $\theta \in T'$ then $gP_i(\theta)\, e$

(ii) if $\theta \notin T'$, then $eR_i(\theta)\, g$.

There are four possibilities, call them I, II, III, IV. If $i = 1$ then either (I) $T' = \{\theta_1, \theta_2\}$ so $gP_1 e$ and $eR_1' g$, or (II) $T' = \{\theta_3, \theta_4\}$ so $gP_1' e$ and $eR_1 g$. Similarly, there are two possibilities (III and IV) for the case $i = 2$.

Consider first possibility I, where $T' = \{\theta_1, \theta_2\}$. Consider the following deception $\delta_i(\theta) = \{\theta_1, \theta_2\}$ if $\theta \in \{\theta_1, \theta_2\}$ and $\delta_i(\theta) = \{\theta\}$ otherwise for all i. Consider the compromise selection where $b(T') = g$. Clearly, δ is incentive compatible with respect to b. Since $g \notin \bigcap_{t \in T'} F(t)$, and F is reliably monotonic, there is some state θ where some agent of some type has an objection. This state cannot be θ_3 or θ_4 as $\delta_{-i}(\theta)$ for $\theta \in \{\theta_3, \theta_4\}$ for all i is a singleton. Therefore as F is reliably monotonic and $g \notin \bigcap_{t \in T'} F(t)$, there is $\theta' \in T'$ and $y \in F(\theta')$ such that:

(i) $yP_2(\theta')\, g$

(ii) if $\theta \in T' \backslash \{\theta'\}$, then $gR_2(\theta)\, y$.

Again there are two possibilities to consider: (Ia) $\theta' = \theta_1$ or (Ib) $\theta' = \theta_2$.

(Ia) If $\theta' = \theta_1$ then $R_2(\theta') = R_2$ and $F(\theta') = a$. From (i) and (ii) it follows that $aP_2\, g$ and $gR_2' a$. Thus, g_2 must be in area G_2 in Fig. 4. Then $g_{21} > \omega_1 - K_1$ and $g_{22} > K_{2'}$, so $g_{11} < K_1$ and $g_{12} < \omega_2 - K_2$. Thus, g_1 belongs to the area

$B_1 \subset C_1$ of Fig. 3. By construction, if an indifference curve for preferences R_1 passes through this area, then it coincides throughout the consumption set with an indifference curve for preferences R'_1. However, this contradicts gP_1e and eR'_1g.

(Ib) This case is completely symmetric to (Ia).

Thus, possibility I leads to a contradiction. The remaining possibilities II, III, IV lead to similar contradictions. Thus, F cannot be reliably monotonic.

6. CONCLUSION

This paper has defined a new notion of interactive implementation and investigated the types of social choice rules that can be interactively implemented. Our analysis suggests that at least the following questions are of interest:

(1) There may be other restrictions on beliefs "off the equilibrium path" worth analyzing.

(2) Since messages in our model are cheap talk, it is necessary to postulate that the planner understands the "language" which the agents speak. On the other hand, if messages were costly to send, standard refinements such as stability could be more powerful. It is clear that making messages costly may well be in the planner's interest.

(3) Allowing for other types of interaction (i.e., having the planner move at the same time as the agents or many times) between the planner and agents may alter the set of social choice rules that can be interactively implemented in an interesting manner—Baliga and Sjöström [3] have made some preliminary investigations along these lines.

(4) The set of social choice rules that can be interactively implemented when there is incomplete information among the agents remains to be characterized.

(5) The principal may be able to commit to an outcome function in some minimal way. For example, the principal may commit not to change the outcome from a to b if the expected gain is smaller than some $\varepsilon > 0$.

(6) Even if the principal cannot commit to an outcome function (for example because messages are unverifiable to third parties), he may be able to commit to a "constitution" which limits his actions, i.e., which restricts the set A from which he can choose. Such a commitment can clearly make the principal better off.

REFERENCES

1. S. Baliga, L. Corchon, and T. Sjöström, "The Theory of Implementation when the Planner is a Player," DAE Working Paper (Economic Theory), No. 9512, Cambridge University, Cambridge, 1995.
2. S. Baliga and T. Sjöström, Interactive implementation, mimeo, Harvard, 1995.
3. S. Baliga and T. Sjöström, work in progress.
4. G. Becker, A theory of social interactions, *J. Polit. Econ.* **82** (1974), 1063–1094.
5. B. Chakravorty, L. Corchon, and S. Wilkie, Credible implementation, *Games Econ. Behavior*, in press.
6. I. K. Cho and D. Kreps, Signalling games and stable equilibria, *Quart. J. Econ.* **102** (1987), 179–221.
7. J. Farrell, Meaning and credibility in cheap-talk games, *Games Econ. Behavior* **5** (1993), 514–31.
8. S. Grossman and M. Perry, Perfect sequential equilibrium, *J. Econ. Theory* **39** (1986), 97–119.
9. E. Maskin, Nash equilibrium and welfare optimality, mimeo, MIT, 1977.
10. E. Maskin and J. Moore, Implementation with renegotiation, mimeo, Harvard, 1988.
11. E. Maskin and J. Tirole, The principal–agent relationship with an informed principal II: Common values, *Econometrica* **60** (1992), 1–42.
12. M. Osborne and A. Rubinstein, "A Course in Game Theory," MIT Press, Cambridge, MA, 1994.
13. D. Ray and K. Ueda, Egalitarianism and incentives, *J. Econ. Theory*, forthcoming.
14. A. K. Sen, Peasants and dualism with or without surplus labor, *J. Polit. Econ.* **74** (1966), 425–50.