Available online at www.sciencedirect.com
ScienceDirect
Games and Economic Behavior 57 (2006) 18–36
GAMES and Economic Behavior
www.elsevier.com/locate/geb

Credible implementation

Bhaskar Chakravorty [a], Luis C. Corchón [b,*], Simon Wilkie [c]

[a] *Monitor Company, Cambridge, MA*
[b] *Department of Economics, Universidad Carlos III, c/ Madrid 126, Getafe, 28903, Madrid*
[c] *California Institute of Technology*

Received 12 September 1993

Available online 20 February 2006

Abstract

The theory of implementation abounds with mechanisms with intricate systems of rewards and punishments off-the-equilibrium path. Generally, it is not in the designer's best interest to go through with the reward/punishment in the "subgame" arising from some disequilibrium play. This would make the mechanism's outcome function non-credible. We define a notion of credible implementation and, in the domain of exchange economies, we show that (a) the non-dictatorial Pareto correspondence can be credibly implemented (b) there is no credibly implementable Pareto-efficient and individually rational social choice rule (SCR) and (c) there is no credibly implementable Pareto-efficient and envy-free SCR. We derive necessary and sufficient conditions for credible implementability of SCR. The main implication is that it is sub-optimal for the designer to be endowed with "too much" information about the economy. Finally, we show that the negative results persist under weaker credibility requirements.

JEL classification: C70; D70

Keywords: Credibility; Implementation; Renegotiation; Commitment

1. Introduction

The literature on implementation abounds with clever mechanisms whose equilibrium outcomes are optimal according to some social choice rule. However, the cleverness of these mechanisms relies on intricate systems of rewards and punishments for deviations from the equilibrium

* Corresponding author. Fax: +34 916249875/9329.
E-mail address: lcorchon@eco.uc3m.es (L.C. Corchón).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2005.09.001

path. This literature generally ignores the designer's incentives to operate the mechanism ex post. These incentives are important if the designer is an interested principal (as in the principal-agent literature) or a government or a body of representatives of the agents whose objectives are embodied in the social choice rule being implemented. In particular, the players must ask themselves if they believe that the planner would enforce outcomes that are undesirable from her point of view. That is, they should ask themselves, is the mechanism "credible"? This paper introduces the concept of "credible implementation" and examines the impact on the implementation problem when we account for the designer's incentives to enforce the outcome selected by the mechanism.

In the standard formulation of the implementation problem, agents, $1, \ldots, n$, are characterized by their preferences, $\succcurlyeq_i$ for each agent i, over a space of allocations, A. The profile $\succcurlyeq = (\succcurlyeq_1, \ldots, \succcurlyeq_n)$ is drawn from a space of potential profiles. In addition, the preferences of the designer are represented by a social choice rule, φ, which picks outcomes in A for every preference profile. The implementation problem is to find a game form, Γ, specifying (i) a message space, S_i for every i, and (ii) an outcome function, g, which maps from $S_1 \times \cdots \times S_n$ into A, so that whenever the true profile is $\succcurlyeq$, the set of equilibrium allocations of the game $(\Gamma, \succcurlyeq)$ coincides with $\varphi(\succcurlyeq)$. The problem we consider is as follows. Assume that a planner has designed a game form Γ that implements φ. Now suppose a disequilibrium message profile, s', is played by the agents; it is not always in the designer's best interest (interpreted in terms of the announced rule φ) to go through with the reward/punishment, $g(s')$, in this "subgame." If the designer does not select according to the stated outcome function g, then the credibility of the mechanism is severely undermined. The assumption implicit in the standard implementation model is that the designer can commit to the mechanism; the commitment presumably is enforced through some "reputation" argument. However, such an argument would indicate that there is a repeated nature to the mechanism design process; such repetition would also alter the equilibria of the mechanism at any given point in time, generally by adding new unwanted equilibria, thereby destroying its implementation properties in the single-stage game. Hence the standard model (which is based on a static mechanism design assumption) itself does not address the issue at hand. There are extreme examples of non-credible mechanisms in the literature: in certain circumstances, some mechanisms confiscate all resources from the agents, or from any given agent; and others promise an agent whatever outcome the agent desires, regardless of the effect on the rest of the economy. In general, the lack of credibility appears in more subtle ways, and in a variety of applications. Consider, for example, the theory of auctions. A credibility problem arises in the literature on optimal auctions, where a minimum reserve price is set and a revenue-maximizing auctioneer must commit to rejecting positive bids and withholding the object on sale if all bids fall below the reserve price. In practice we often see sellers privately contracting with buyers after the auction closes which undermines the role of the reserve price.[1]

In this paper, we address the credibility issue in the context of Nash implementation. We define a notion of "credible implementation," that roughly speaking, requires that the allocations selected by the mechanism always lie in the range of the social choice rule. That is, the planner can justify ex post the chosen outcome. This is not the only possible way to model "credibility" in this setting, but we feel that it captures in a simple way some essential elements of the problem. We derive necessary and sufficient conditions for credible implementation of a social choice rule and show that for exchange economies:

[1] We thank a referee for this suggestion.

(a) the non-dictatorial Pareto correspondence can be credibly implemented,
(b) there exists no Pareto-efficient and individually rational social choice rule that can be implemented in a credible manner, and
(c) there exists no envy-free and Pareto-efficient social choice rule that can be implemented in a credible manner.

Next, we address the following question: are these negative results an outcome of an excessively strong credibility requirement? We explore this issue with respect to the result (b) above. In defining a weaker requirement, that we call *weak credibility*, we impose more structure on the standard implementation problem. We require not only that the target social choice rule be common knowledge, but the entire social utility function that the designer is maximizing needs to be specified. Note that the social choice rule simply conveys information about the allocations that maximize social utility for each preference profile. We show that for any social utility function satisfying a weak property of unanimity, the negative results persist: there exists no Pareto-efficient and strictly individually rational social choice rule that can be implemented in a weakly credible manner. A similar result is obtained if the social utility function were to satisfy another weak property of "limited preference for efficiency." This indicates that the negative implications of our notion of credibility are rather robust.

The central implication of our paper is that it is sub-optimal for the designer to have "too much" information about the economy. We shall argue that the designer's freedom to design credible mechanisms decreases as more information is observable regarding the economy. This is because the more information in the hands of the planner, the less allocations she can use in a credible manner when designing the mechanism. Our conditions show just how much absence of information is sufficient to ensure credibility of implementation of sub-correspondences of the social choice rule. Of course, there is a "discontinuity" at the limit as we proceed towards increasing informativeness on the part of the designer: when the latter has complete information, the mechanism design problem itself disappears.

We close this section with two comments on the concept introduced in this paper.

1.1. Credibility and sequential equilibrium

In spirit, the notion of credible implementation is akin to the idea of a sequential equilibrium of a game between economic agents as first movers and the social planner as a second mover. The social planner is modeled as a dummy player whose preferences are embodied in the social choice rule. The agents simultaneously transmit messages to the designer who chooses an outcome in response to the messages received. For a mechanism to be credible, we require the outcome chosen to be a "best response" (consistent with knowledge of the social choice rule) according to some posterior beliefs about the underlying economy. On the other hand, for a mechanism to be weakly credible, we require the outcome chosen to be a "best response" (consistent with knowledge of the social utility function) according to some posterior beliefs about the underlying economy. Hence, our notion of implementation incorporates the idea of a sequential equilibrium, but is not one in the formal sense. Baliga et al. (1997) and Baliga and Sjostrom (1999) assume that the planner is a full-fledged player, able to anticipate the strategies played by agents and to update priors. With respect to the approach presented here it has the advantage of modeling incentives of the planner more in line with standard game theory but it has the drawback of requiring a refinement of perfect Bayesian equilibrium that, in certain contexts, may be debatable. Also this approach does not produce clear cut results.

1.2. Related literature

The questions associated with the inability to commit to a mechanism have been addressed in Maskin and Moore (1999), Rubinstein and Wolinsky (1992), Aghion et al. (1994), Ray and Ueda (1996) and Jackson and Palfrey (2001). The approach taken in these papers is that there is some renegotiation or bargaining process after the outcome is chosen by the planner, which leads to efficiency ex post. Our approach is different. Instead of focusing on the issue of lack of control on the actions of the agents (moral hazard), we consider that of commitment on the part of the designer. Thus, we do not extend the original model to allow for ex post renegotiation possibilities. Instead, the designer's objectives are consistently associated with the social choice rule she is committed to implement.

2. Preliminaries

Let us formally introduce the problem. The set of agents is N. Let i be a typical element of N. Let A denote the set of feasible allocations. Let $\mathcal{R}_i$ denote the domain of admissible preference relations for i, defined on $A \times A$. The set $\mathcal{R}$ completely characterizes the class of economies under consideration, and we refer to $\succcurlyeq \in \mathcal{R}$ as an *economy*. To simplify the exposition, we assume that $\mathcal{R}$ is finite.[2]

Let $L_i(z, \succcurlyeq) \equiv \{z' \in A\colon z \succcurlyeq_i z'\}$ denote the (weak) lower contour set for i defined at allocation z in the economy $\succcurlyeq$. A *social choice rule* is a non-empty valued correspondence $\varphi : \mathcal{R} \to\!\!\to A$.

A *game form* Γ is a pair $\langle S, g\rangle$, where S_i is a message space for agent i, $S = \times_{i\in N} S_i$ and $g : S \to A$ is an outcome function. A strategy for i in Γ is a function $\sigma_i : \mathcal{R} \to S_i$, with Σ_i denoting the strategy space.[3]

The set of *Nash equilibria of* Γ is denoted by $NE(\Gamma) \equiv \{\sigma \in \Sigma\colon \forall i \in N,\ \forall s_i \in S_i,\ \forall \succcurlyeq \in \mathcal{R}, g(s_i, \sigma_{-i}(\succcurlyeq)) \in L_i(g(\sigma_i(\succcurlyeq), \sigma_{-i}(\succcurlyeq), \succcurlyeq)\}$.

The set of *Nash equilibrium allocations of* Γ *in* A is denoted by $NE_A(\Gamma, \succcurlyeq) \equiv \{z \in A\colon \exists \sigma \in NE(\Gamma), g(\sigma(\succcurlyeq)) = z\}$.

Given a prior β^o a social choice rule φ is *Nash implementable* if there is a game form such that for each economy in the domain the set of Nash equilibria yield allocations that coincide with those prescribed by φ.

The social planner's objective is to implement φ. The planner cannot observe the true economy $\succcurlyeq$ and possesses a prior probability distribution defined on the domain of economies, denoted $\beta^o : \mathcal{R} \to [0, 1]$. The distribution β^o is common knowledge and its support is denoted by $supp(\beta^o)$. Let $\mathcal{B}$ be the class of admissible priors. If $supp(\beta^o)$ is a singleton, then the designer has perfect knowledge of the true economy; once there are two or more elements in $supp(\beta^o)$, the implementation of φ becomes a non-trivial issue.

Given a planning problem, the agents and the designer participate in the following process. The designer chooses Γ; then the agents (simultaneously) submit their messages in Γ to the designer, who in turn decides on an allocation by applying the outcome function in Γ.

[2] This is because subsequently, we shall define probability distributions on $\mathcal{R}$. The finiteness assumption allows us to make the main points without measure-theoretic machinery.

[3] σ_i is defined on $\mathcal{R}$ and not on $\mathcal{R}_i$ reflecting the fact that there is complete information among the agents. This is because we use Nash equilibrium as our solution concept.

Next, we introduce the criterion central to this paper, for which some additional notation is needed: Let $Im(g) \equiv \{g(s) \in A\colon\ s \in S\}$ and

$$\tilde{\varphi}\big(\beta^o\big) = \varphi\big(supp\big(\beta^o\big)\big) \equiv \big\{z \in \varphi(\succcurlyeq) :\succcurlyeq \in supp\big(\beta^o\big)\big\}.$$

Definition 1. A social choice rule φ is *credibly implementable in* $\mathcal{B}$ if $\forall \beta^o \in \mathcal{B}$, $\exists \Gamma = \langle S, g\rangle$ such that

(i) $\forall \succcurlyeq \in supp(\beta^o), NE_A(\Gamma, \succcurlyeq) = \varphi(\succcurlyeq)$, and
(ii) $Im(g) \subseteq \varphi(supp(\beta^o))$.

The definition is in two parts. We are interested in "global" implementability as in Palfrey and Srivastava (1987) and Chakravorti (1992, 1993), i.e. the implementability of φ in every planning problem. Since each problem corresponds to a prior beliefs distribution, both conditions must hold for any prior beliefs that the planner may have about the economy. Condition (i) requires that the planner must design a game form whose equilibrium allocations coincide with the φ-optimal allocations for any economy considered possible according to the beliefs β^o. Condition (ii) says that the game form must select outcomes that are φ-optimal for *some* economy considered possible by the beliefs β^o. The latter condition is the credibility restriction: the planner must be able to rationalize any outcome she promises by arguing that the outcome is optimal for some economy that she considers to be possible. This may be thought of a planner who is highly intolerant of picking the wrong outcome. However, at the end of the day a decision must be implemented so if the price to the planner of selecting the wrong allocation is sufficiently high, then the planner would never pick an outcome that can never be optimal. We discuss some alternative notions that capture the idea of credibility in the final section.

Next, we shall define some properties that are used in the analysis of credible implementability of the rules just defined.

Definition 2. A social choice rule φ satisfies *Monotonicity Relative to* $\mathcal{B}$ ($MON\mathcal{B}$) if $\forall \beta^o \in \mathcal{B}$, $\forall \succcurlyeq, \succcurlyeq' \in \sup(\beta^o)$,

$$\big[z \in \varphi(\succcurlyeq) \text{ and } \big\{L_i(z, \succcurlyeq) \cap \tilde{\varphi}\big(\beta^o\big)\big\} \subseteq L_i(z, \succcurlyeq')\big] \quad \Rightarrow \quad \big[z \in \varphi(\succcurlyeq')\big].$$

When $\tilde{\varphi}(\beta^o) = A$ then $MON\mathcal{B}$ reduces to the familiar notion of "Maskin-monotonicity," due to Maskin (1999), which is given in the next definition. The usual definition, is modified to apply globally in $\mathcal{B}$.

Definition 3. A social choice rule φ satisfies *Maskin-monotonicity* if $\forall \succcurlyeq, \succcurlyeq' \in \sup(\beta^o)$,

$$\big[z \in \varphi(\succcurlyeq) \text{ and } L_i(z, \succcurlyeq) \subseteq L_i(z, \succcurlyeq')\big] \quad \Rightarrow \quad \big[z \in \varphi(\succcurlyeq')\big].$$

Definition 4. A social choice rule φ satisfies *No Veto Power Relative to* β^o ($NVP\beta^o$) if $\forall \succcurlyeq \in \sup(\beta^o)$,

$$\big[\exists z \in \tilde{\varphi}\big(\beta^o\big) \text{ and } i \in N \text{ such that } \forall z' \in \tilde{\varphi}\big(\beta^o\big),\ \forall j \in N \setminus \{i\},\ z_j \succcurlyeq_j z'\big] \quad \Rightarrow \quad \big[z \in \varphi(\succcurlyeq)\big].$$

When $\tilde{\varphi}(\beta^o) = A$, then NVP$\beta^o$ is the "No Veto Power" condition of Maskin (1999), in which the allocation z in the definition above is $\succcurlyeq$-maximal in A for $n-1$ agents. Our condition simply says that if in a given economy, $n-1$ agents agree on an allocation as top ranked among all

allocations in $\tilde{\varphi}(\beta^o)$, then that allocation must be φ-optimal for the economy. Note that Maskin's No Veto Power condition is satisfied by any φ in the class of exchange economies. However, NVPβ^o is not necessarily met unless there is conflict of interests between each pair of agents. Maskin (1985) shows that if $n > 2$ in "economic" environments, such as those considered here, Maskin-monotonicity is equivalent to Nash implementability.

Next, we define some key social choice rules whose implementability is considered below: The *Pareto-efficiency* social choice rule, *PE* is defined by

$$PE(\succcurlyeq) = \{z \in A\colon \nexists z' \in A \text{ such that } \forall i \in N,\ z' \succcurlyeq_i z \text{ and for some } i, z' \succ_i z\}.$$

The *non-dictatorial Pareto-efficiency* social choice PE^* is defined by

$$PE^*(\succcurlyeq) = \{z \in PE(\succcurlyeq)\colon \forall i \in N,\ z_i \neq 0\}.$$

The *individual rationality* social choice rule, *IR*, is defined by

$$IR(\succcurlyeq) = \{z \in A\colon \forall i \in N,\ z_i \succcurlyeq_i \omega_i\}.$$

The *envy-freeness* social choice rule *EF* is defined by

$$EF(\succcurlyeq) = \left\{z \in A\colon \forall i \in N,\ \nexists j \in N \setminus \{i\},\ z_j \succ_i z_i\right\}.$$

3. Results on credible implementation

In this section, we establish necessary and sufficient conditions for credible implementability of social choice correspondences. Furthermore, we identify correspondences that are credibly implementable and those that are not. The latter includes some correspondences that are Nash-implementable. Finally, we establish conditions on the designer's beliefs that ensure credible (partial) implementability of a large class of Nash-implementable correspondences. By partial implementability, we mean that for any φ, there is a sub-correspondence $\varphi' \subseteq \varphi$ which is implementable. Given that the designer is, typically, indifferent over the allocations in $\varphi(\succcurlyeq)$ for given $\succcurlyeq$, partial implementation of φ is quite acceptable.

Theorem 1. *If φ is credibly implementable in some $\mathcal{B}$, then it satisfies MON$\mathcal{B}$.*

Proof. The standard proof (see, for example, McKelvey, 1989) showing that Maskin-monotonicity is necessary for Nash implementability may be adapted to prove this theorem by taking $\tilde{\varphi}(\beta^o)$ instead of A as the set of alternatives. □

Theorem 2. *Suppose $n > 2$. If φ satisfies MON$\mathcal{B}$ and NVPβ^o for all $\beta^o \in \mathcal{B}$, then φ is credibly implementable in $\mathcal{B}$.*

Proof. The proof of this theorem is also readily obtained through the standard constructive techniques (see, e.g. McKelvey, 1989) by replacing A with $\tilde{\varphi}(\beta^o)$. □

The standard constructive proofs for sufficiency of Nash implementability rely on the ability of agents to obtain their most desired outcome in certain parts of the game (such as the one that corresponds to the chasing of the highest number in McKelvey, 1989) and thereby reducing the set of equilibrium outcomes. The credibility constraint restricts what is achievable in this part of the game.

In order to shed some light on the concept of credible implementation we focus our attention on the domain of exchange economies. First, let us briefly review the results on Nash implementation in exchange economies. Hurwicz (1979) established that the (constrained) Walrasian correspondence is implementable. Then, Maskin (1985) showed that the Pareto-efficient and individually rational correspondence is implementable. Later on Thomson (1987) proved that the Pareto-efficient and envy-free correspondence is implementable. It is easily shown that the non dictatorial Pareto correspondence is also implementable.

The following theorems identify social choice rules that meet Property 1, and those that do not. We first study the consequences of requiring efficiency and focus our attention on the PE^* social choice rule. Later on we will consider social choice rules that select efficient allocations fulfilling additional criteria like the $PE \cap IR$ and $PE \cap EF$ correspondences.

Theorem 3. *The correspondence PE^* satisfies $MON\mathcal{B}$ for any $\mathcal{B}$.*

Proof. Choose $\succcurlyeq, \succcurlyeq' \in \mathcal{R}$ and $\beta^o \in \mathcal{B}$ with $\succcurlyeq, \succcurlyeq' \in supp(\beta^o)$. Suppose that $z \in PE^*(\succcurlyeq)$. Also suppose that the hypothesis of $MON\mathcal{B}$ is met, i.e. for all $i \in N$, $\{L_i(z, \succcurlyeq) \cap \widetilde{PE}^*(\beta^o)\} \subseteq L_i(z, \succcurlyeq')$, but $z \notin PE^*(\succcurlyeq')$. Since $z \notin PE^*(\succcurlyeq')$, there exists $z'' \in A$ such that z'' Pareto-dominates z in the economy $\succcurlyeq'$. By transitivity and closedness of preferences, the Pareto-dominance relation is transitive and closed. Thus, as A is compact (we are in an exchange economy), there exists $z^* \in PE(\succcurlyeq')$ that Pareto-dominates z in the economy $\succcurlyeq'$. By hypothesis, $z_i \neq 0, \forall i \in N$. Thus by monotonicity of preferences, $z_i^* \neq 0, \ \forall i \in N$. Thus, $z^* \in PE^*(\succcurlyeq')$, and therefore $z^* \in \widetilde{PE}^*(\beta^o)$. However, by assumption, $\{L_i(z, \succcurlyeq) \cap \widetilde{PE}^*(\beta^o)\} \subseteq L_i(z, \succcurlyeq'), \ \forall i \in N$. Thus $\forall i \in N, z^* \succcurlyeq_i z$. But this contradicts the assumption that $z \in PE^*(\succcurlyeq)$. □

In the two-person case, the intuition underlying the argument is illustrated in Fig. 1. Suppose that the hypothesis of $MON\mathcal{B}$ is met and its conclusion is not. Without loss of generality, choose $z \in PE^*(\succcurlyeq)$ as the point where the indifference curves for agent 1 in the two economies $\succcurlyeq$ and $\succcurlyeq'$ intersect. Such an intersection must occur to ensure that $z \notin PE^*(\succcurlyeq')$. Next, consider the intersection of the set $PE^*(\succcurlyeq')$ with the indifference curve for agent 1 through z in economy $\succcurlyeq'$. This intersection must occur to the south-east of z, otherwise we would violate the requirement that for all $i \in N$, $\{L_i(z, \succcurlyeq) \cap \widetilde{PE}^*(\beta^o)\} \subseteq L_i(z, \succcurlyeq')$. Let z^* be such a point of intersection. Now observe that $z^* \in \widetilde{PE}^*(\beta^o)$ and is strictly preferred to z by agent 2 in economy $\succcurlyeq'$. However, if $z \in PE^*(\succcurlyeq)$, agent 2 must clearly prefer z to z^* in economy $\succcurlyeq$. This violates the hypothesis that $\{L_2(z, \succcurlyeq) \cap \widetilde{PE}^*(\beta^o)\} \subseteq L_2(z, \succcurlyeq')$.

Corollary to Theorem 3. *Suppose $n > 2$. The PE^* correspondence is credibly implementable for any $\mathcal{B}$.*

Proof. The PE^* correspondence satisfies $NVP\beta^o$ trivially for every β^o since no agent has a $\succcurlyeq$-maximal element in $\widetilde{PE}^*(\beta^o)$ for any $\succcurlyeq$. By Theorems 2 and 3, PE^* is credibly implementable in $\mathcal{B}$. □

Unfortunately our two next results show that if, in addition to Pareto-efficiency, individual rationality or envy-freeness are required, we obtain negative results. The difference between this (negative) results and those obtained in Theorem 3 (positive results) is explained by the fact that the addition of an extra requirement reduces the set of allocations that the planner can choose from and this makes, in some cases, implementation impossible.

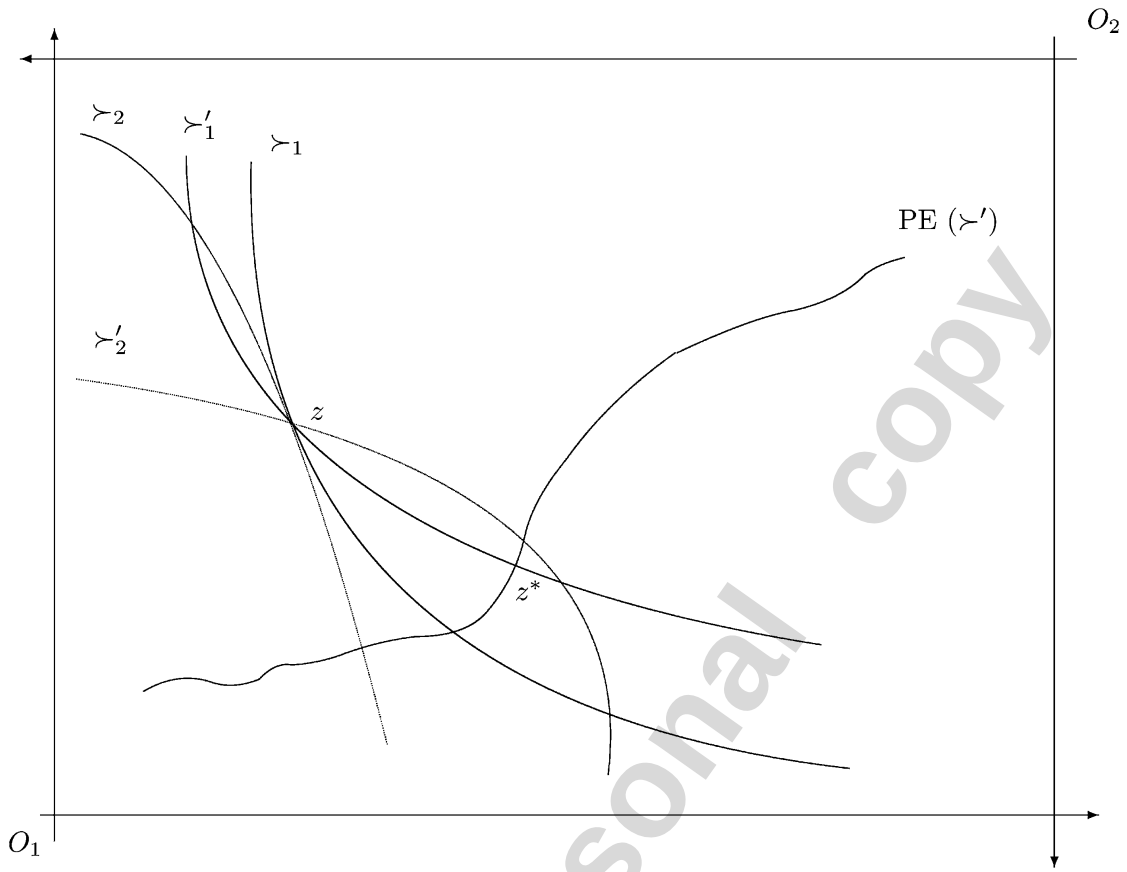

Fig. 1.

Theorem 4. *Suppose $\varphi \subseteq PE \cap IR$. Then φ fails to satisfy MON$\mathcal{B}$ for some $\mathcal{B}$.*

Proof. Consider $\beta^o : \mathcal{R} \to [0, 1]$ such that $\beta^o(\succcurlyeq) = \beta^o(\succcurlyeq') = 0.5$ and the $PE \cap IR$ correspondence defined for $\succcurlyeq$ and $\succcurlyeq'$ is as given in Fig. 2. $PE \cap IR(\succcurlyeq) = AB$ and $PE \cap IR(\succcurlyeq') = CD$. The example is for a two-person, two-good economy. However, it easily generalizes. Choose $z \in \varphi(\succcurlyeq) \subseteq PE \cap IR(\succcurlyeq)$. Observe that $L_1(z, \succcurlyeq) \cap \tilde{\varphi}(\beta^o) \subseteq CD \cup AE$ and $L_2(z, \succcurlyeq) \cap \tilde{\varphi}(\beta^o) \subseteq EB$. By construction, $CD \cup AE \subseteq L_1(z, \succcurlyeq')$ and $EB \subseteq L_2(z, \succcurlyeq')$. Clearly, $z \notin PE \cap IR(\succcurlyeq')$. □

For economies for which initial endowments are collectively owned, the concept of individual rationality is replaced by the concept of an envy-free allocation.

Theorem 5. *Suppose $\varphi \subseteq PE \cap EF$. Then φ fails to satisfy MON$\mathcal{B}$ for some $\mathcal{B}$.*

Proof. The argument is identical to that of the previous proof. Consider $\beta^o : \mathcal{R} \to [0, 1]$ such that $\beta^o(\succcurlyeq) = \beta^o(\succcurlyeq') = 0.5$ and the $PE \cap EF$ correspondence defined for $\succcurlyeq$ and $\succcurlyeq'$ is as given in Fig. 3. The figure is drawn so that the vectors $O_1\vec{X} = O_2\vec{Y}$ and $O_2\vec{X} = O_1\vec{Y}$. The locations of the points A, B, C and D are chosen so that $PE \cap EF(\succcurlyeq) = AB$ and $PE \cap EF(\succcurlyeq') = CD$. □

The intuition underlying the examples is as follows. Suppose the planner were "reasonably well" informed, but not fully informed about the true economy. In this case, the support of her priors β^o is very narrow. Consequently, the set of φ-optimal allocations is rather small. This severely restricts the range of the outcome function that the designer can use and, consequently,

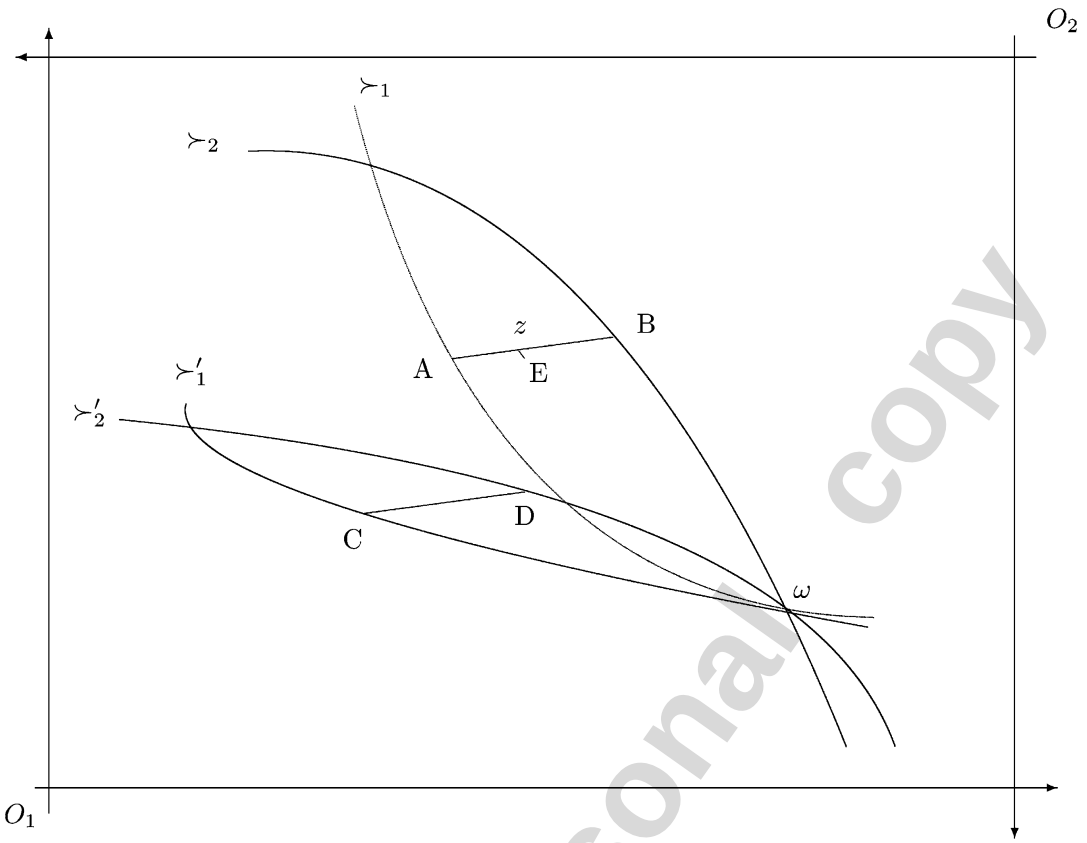

Fig. 2.

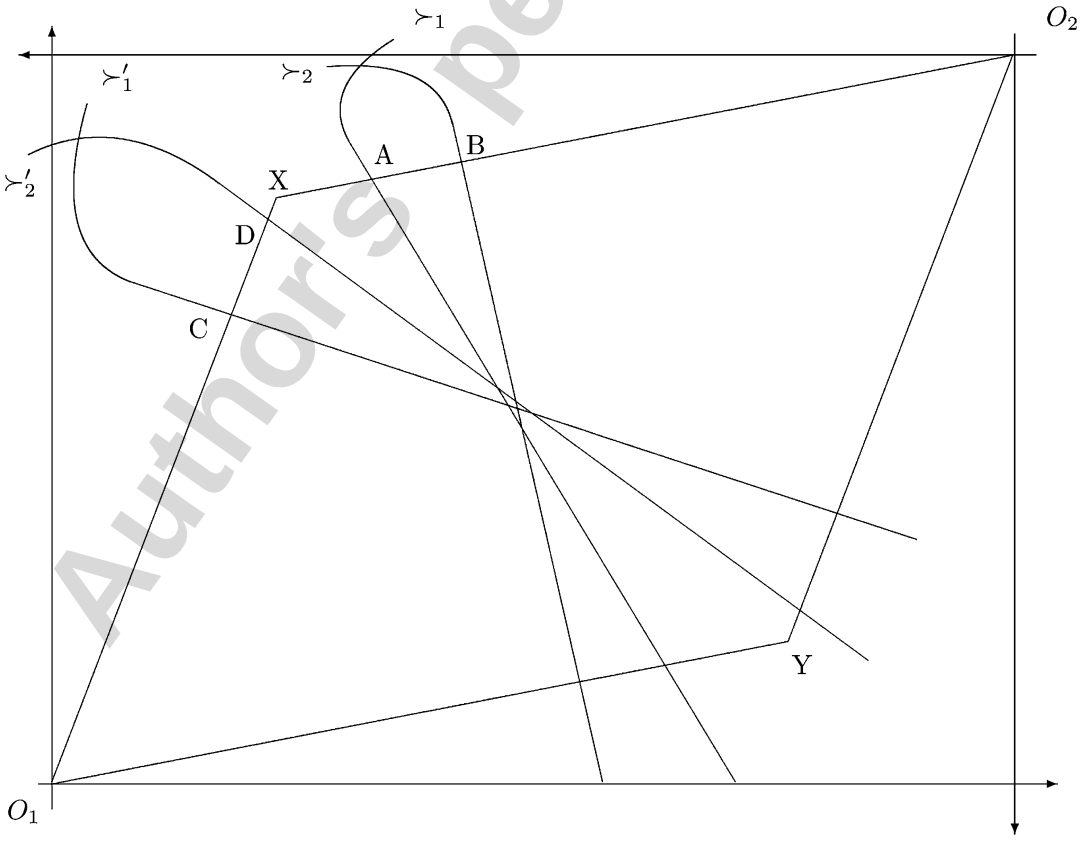

Fig. 3.

there are cases where a single game form satisfying these restrictions cannot be constructed to credibly implement φ. In the examples, the designer is limited by the narrow support for β^o. It is, of course, true that another planner with a sufficiently wide support for her priors could overcome these restrictions. However, we are interested in the (credible) implementability of φ, not with respect to a particular planning problem, but for *any* planning problem: φ should be implementable in every situation regardless of the state of the planner's prior information. It should be noted that the same arguments can easily be shown to hold for economies with at least three people in which $NVP\beta^o$ is satisfied. Hence, the non-implementability of the choice rules hinges entirely on the failure to satisfy $MON\mathcal{B}$. The examples also point that when the designer's information about the economy is relatively good, it is more difficult for her to design a mechanism in a credible manner. This follows from the requirement that the range of the outcome function must be contained in the set of allocations that are φ-optimal for some economy in the support of the designer's priors. Good information means that the support is narrow; hence there is relatively little flexibility. This raises the obvious question: how much incompleteness of information is "enough" to guarantee credible implementability of Nash implementable rules? In order to answer this question we will need the following definitions.

Given Δ as the $(m-1)$-dimensional unit simplex, and $p \in \Delta$, $\forall i \in N$ define $\overline{H}_i(z_i, p) = \{z' \in A,\ pz' = pz_i\}$ and $H_i(z_i, p) = \{z' \in A\colon\ pz \leqslant pz_i\}$. Given $p \in \Delta$, let $\succcurlyeq^p \in \mathcal{R}$ be defined by

$$\forall i \in N,\ \forall z \in A,\ \left\{z' \in A\colon\ z \succcurlyeq_i^p z'\right\} = H(z, p).$$

These are the preferences with linear indifference surfaces with normal p. Then, the (*constrained*) *Walrasian* social choice rule, W, is defined by

$$W(\succcurlyeq) = \left\{z \in A\colon\ \exists p \in \Delta,\ \forall i \in N,\ H_i(\omega, p) \subseteq L_i(z, \succcurlyeq)\right\}.$$

Also, define $z^e \in A$ by $z_i^e = z_j^e,\ \forall i, j \in N$. Then, the (*constrained*) *Walrasian from equal-division* social choice rule, W^e, is defined by

$$W^e(\succcurlyeq) = \left\{z \in A\colon\ \exists p \in \Delta,\ \forall i \in N,\ H_i\left(z^e, p\right) \subseteq L_i(z, \succcurlyeq)\right\}.$$

Let $\Delta^W(\succcurlyeq)$ and $\Delta^=(\succcurlyeq)$ denote, respectively, the sets of Walrasian prices and Walrasian prices from equal-division for the economy $\succcurlyeq$. The following properties are conditions on the designer's beliefs.

Definition 5. A prior distribution $\beta^o : \mathcal{R} \to [0, 1]$ satisfies the *Linear Preferences Property* (*LPP*) if

$$\exists p \in \Delta \text{ and } \succcurlyeq^p \in \mathcal{R} \text{ such that } \succcurlyeq^p \in supp(\beta^o).$$

Definition 6. A prior distribution $\beta^o : \mathcal{R} \to [0, 1]$ satisfies $LPP(W)$ if

$$\left[\succcurlyeq \in supp(\beta^o)\right] \quad \Rightarrow \quad \left[\exists p \in \Delta^W(\succcurlyeq) \text{ and } \succcurlyeq^p \in \mathcal{R} \text{ such that } \succcurlyeq^p \in supp\left(\beta^o\right)\right].$$

Definition 7. A prior distribution $\beta^o : \mathcal{R} \to [0, 1]$ satisfies $LPP(=)$ if

$$\left[\succcurlyeq \in supp(\beta^o)\right] \quad \Rightarrow \quad \left[\exists p \in \Delta^=(\succcurlyeq) \text{ and } \succcurlyeq^p \in \mathcal{R} \text{ such that } \succcurlyeq^p \in supp\left(\beta^o\right)\right].$$

The restrictions on the beliefs given in the definitions above have the following form: the designer's priors must assign positive probability to at least one "linear" economy in the domain of possible economies. $LPP(W)$ and $LPP(=)$ distinguish two criteria that such a linear economy

must satisfy; if an economy $\succcurlyeq$ is admissible then the priors must admit linear economies defined by a Walrasian price for $\succcurlyeq$ (in the case of $LPP(W)$) and a Walrasian price from equal-division for $\succcurlyeq$ (in the case of $LPP(=)$).

Definition 8. A social choice rule φ is *Pareto-indifferent* if

$$\big[z \in \varphi(\succcurlyeq) \text{ and } \forall i \in N,\ z_i' \sim_i z\big] \quad \Rightarrow \quad \big[z' \in \varphi(\succcurlyeq)\big].$$

This property is from Gevers (1986). It is a natural condition which is met by most choice rules of interest. The property is closely related to two regularity assumptions in Thomson (1984, 1987).[4]

The following result establishes that $LPP(W)$, applicable everywhere in $\mathcal{B}$, together with the mild Pareto-indifference assumption implies that any Nash-implementable Pareto-efficient and individual rational rule can be partially credibly implemented in the sense that there is a subcorrespondence of such a rule that is implementable, provided that the (constrained) Walrasian correspondence, W, satisfies $NVP\beta^o$ in every β^o. The role of the W correspondence in the credible implementation of a social choice rule becomes evident upon examination of the proof of the theorem.

Theorem 6. *Suppose $n > 2$. If*

(i) $\varphi \subseteq PE \cap IR$,
(ii) φ *satisfies Pareto-indifference,*
(iii) φ *satisfies Maskin Monotonicity,*
(iv) *every $\beta^o \in \mathcal{B}$ satisfies $LPP(W)$,*
(v) *W satisfies $NVP\beta^o$ for every $\beta^o \in \mathcal{B}$,*

then there exists $\varphi' \subseteq \varphi$ such that φ' is credibly implementable in $\mathcal{B}$.

Proof. First, in Lemma 1 below we show that if every $\beta^o \in \mathcal{B}$ satisfies $LPP(W)$, then the (constrained) Walrasian correspondence W satisfies $MON\mathcal{B}$. Therefore, as condition (v) holds, by Theorem 2, W is credibly implementable in $\mathcal{B}$. Lemma 2 below shows that if φ is Nash implementable, φ contains W. Thus by condition (iii) and the definition of partial credible implementation the proof is complete. □

Lemma 1. *If every $\beta^o \in \mathcal{B}$ satisfies $LPP(W)$, then W satisfies $MON\mathcal{B}$.*

Proof. Choose $\beta^o : \mathcal{R} \to [0, 1]$ such that $LPP(W)$ is met and choose $\succcurlyeq, \succcurlyeq' \in supp(\beta^o)$ such that $\{L_i(z, \succcurlyeq) \cap \widetilde{W}(\beta^o)\} \subseteq L_i(z, \succcurlyeq')$ for some $z \in W(\succcurlyeq)$. By $LPP(W)$, for some price vector $p \in \Delta$ for which z is $\succcurlyeq$-maximal on $H_i(z, p)$ for all $i \in N$, there exists $\succcurlyeq^p \in supp(\beta^o)$ such that $L_i(z, \succcurlyeq^p) = H_i(z, p)$ for all $i \in N$. By definition of W, $\overline{H}(z, p) \subseteq W(\succcurlyeq^p) \subseteq \widetilde{W}(\beta^o)$. Then, $\{L_i(z, \succcurlyeq) \cap \widetilde{W}(\beta^o)\} \subseteq L_i(z, \succcurlyeq')$ implies that $\overline{H}(z, p) \subseteq L_i(z, \succcurlyeq')$. By definition of W, $z \in W(\succcurlyeq')$. □

[4] Even though we shall subsequently appeal to Thomson's results, we use the Pareto-indifference property since it is sufficient for these results and economizes on the number of new conditions that need to be defined to complete the arguments.

Lemma 2. *If $\varphi \subseteq PE \cap IR$, φ satisfies Maskin-monotonicity and Pareto-indifference, then $W \subseteq \varphi$.*

Proof. It follows from the following adaptation of Thomson's (1984) result.[5] □

The next result establishes a similar set of sufficiency conditions for credible implementability of Nash-implementable rules satisfying Pareto-efficiency and envy-freeness.

Theorem 7. *Suppose $n > 2$. If*

(i) $\varphi \subseteq PE \cap EF$,
(ii) *φ satisfies Pareto-indifference,*
(iii) *φ satisfies Nash implementability,*
(iv) *every $\beta^o \in \mathcal{B}$ satisfies $LPP(=)$,*
(v) *W^e satisfies $NVP\beta^o$ for every $\beta^o \in \mathcal{B}$,*

then there exists $\varphi' \subseteq \varphi$ such that φ' is credibly implementable in $\mathcal{B}$.

Proof. The argument is similar to that needed to establish Theorem 6 using the lemma below. This lemma is taken from Thomson (1987). □

Lemma 3. *If $\varphi \subseteq PE \cap EF$ satisfies Maskin-monotonicity and Pareto-indifference, then $W^e \subseteq \varphi$.*

The results of this section are, in general, negative in nature. Many interesting social choice rules are not credibly implementable. To ensure credibility, the designer's beliefs must be sufficiently diffuse which in turn assures the designer greater flexibility in designing the range of the outcome function. The conditions on the beliefs given above identify sufficient conditions on the extent of information asymmetry that must exist between the agents and the designer to ensure credible implementability. It may be checked that a weaker condition such as *LPP* cannot be used to replace $LPP(W)$ or $LPP(=)$ in the theorems above. Examples can be easily constructed showing that $MON\mathcal{B}$ is violated by any $\varphi \subseteq [PE \cap IR]$ or $\varphi \subseteq [PE \cap EF]$ even if *LPP* is met.

4. Results on weakly credible implementation

The negative results of the previous section are obtained under the assumption that the only information available about the designer's objective is that she is committed to implementing a social choice rule φ. This yields a rather strong credibility requirement. In many situations, of course, there is more information on the designer's objectives; for example, the social utility function may be known in addition to φ. This yields the weaker requirement given in Definition 1B below. In order to highlight the difference with the notion of credible implementation we introduce some extra machinery.

For a given φ, we can define a *social utility function* $u : A \times \mathcal{R} \to \Re$ which is consistent with φ. For each $\succcurlyeq \in \mathcal{R}$, the set of utility-maximizing allocations in A is $\varphi(\succcurlyeq)$. The function u is such that for all $\succcurlyeq \in \mathcal{R}$, for all $z \in \varphi(\succcurlyeq)$, $u(z, \succcurlyeq) \equiv u^{\max}$. The class of *social utility functions consistent*

[5] Thomson uses a regularity condition instead of Pareto-indifference. Under the remaining conditions of Lemma 2, Pareto-indifference implies his condition. An analogous comment holds for Lemma 3 below.

with $\varphi, \mathcal{U}^{\varphi}$, is defined by $\mathcal{U}^{\varphi} = \{u\colon \forall \succcurlyeq \in \mathcal{R}, \forall z \in Z,\ z \in \varphi(\succcurlyeq) \Rightarrow u(z, \succcurlyeq) = \max_{x \in A} u(x, \succcurlyeq\}$. Given $u \in \mathcal{U}^{\varphi}$ and a posterior distribution conditioned on the agents' messages, $\beta : \mathcal{R} \times S \to [0, 1]$, the designer's expected utility from an allocation $z \in A$ is denoted $\tilde{u}(z, \beta)$. In order to see the difference with our previous notion, let us write the definition of credible implementation in a slightly different—but logically equivalent—form.

Definition 1A. A social choice rule φ is *credibly implementable in* $(\mathcal{B}, u)$ if $\forall \beta^o \in \mathcal{B}$, $\exists \Gamma = \langle S, g\rangle$, $\exists \beta : \mathcal{R} \times S \to [0, 1]$ such that $\forall u \in \mathcal{U}^{\varphi}$,

(i) $\forall \succcurlyeq \in \mathit{supp}(\beta^o), NE_A(\Gamma, \succcurlyeq) = \varphi(\succcurlyeq)$,
(ii) $\forall s \in S, [z = g(s)] \Rightarrow [\forall z' \in A,\ \tilde{u}(z, \beta(\cdot, s)) \geqslant \tilde{u}(z', \beta(\cdot, s))]$, and
(iii) $\forall \succcurlyeq \notin \mathit{supp}(\beta^o),\ \forall s \in S, \beta(\succcurlyeq, s) = 0$.

Condition (i) in the definition above is identical to its counterpart in Definition 1. Condition (ii) requires that given any message s, there exist beliefs for the designer conditional on s such that selecting $g(s)$ is a best-response, regardless of whether or not the outcome is expected to be realized in equilibrium.[6] Condition (iii) requires that the conditional beliefs assign probability zero to economies that had zero probability in the prior distribution. We are now prepared to give the definition of weak credible implementation:

Definition 1B. Suppose $u \in \mathcal{U}^{\varphi}$ is common knowledge. A social choice rule φ is *weakly credibly implementable in* $(\mathcal{B}, u)$ if $\forall \beta^o \in \mathcal{B}$, $\exists \Gamma = \langle N, S, g\rangle$, $\exists \beta : \mathcal{R} \times S \to [0, 1]$ such that $\forall \succcurlyeq \in \mathcal{R}$,

(i) $\forall \succcurlyeq \in \mathit{supp}(\beta^o), NE_A(\Gamma, \succcurlyeq) = \varphi(\succcurlyeq)$,
(ii) $\forall s \in S, [z = g(s)] \Rightarrow [\forall z' \in A, \tilde{u}(z, \beta(\cdot, s)) \geqslant \tilde{u}(z', \beta(\cdot, s))]$, and
(iii) $\forall \succcurlyeq \notin \mathit{supp}(\beta^o), \forall s \in S, \beta(\succcurlyeq, s) = 0$.

The conditions of Definition 1B are identical to those of Definition 1A, with the exception that while condition (ii) must hold for *every* $u \in \mathcal{U}^{\varphi}$ in Definition 1A, the corresponding condition in 1B must hold only for the fixed known u. In the case where only φ is known, the designer's choice of outcome $g(s)$ given a list of messages s must be justifiable as a φ-optimal response for some economy she believes to be possible; hence, the condition (ii) of Definition 1A is independent of the social utility function underlying φ. Once u is also common knowledge, this condition needs to hold only with respect to the known u. The latter imposes a less severe restriction on the range of outcomes that g can take.

In this section, we shall suppose that the social utility function is drawn from some natural sub-classes of social utility functions. We shall first define a class of functions characterized by a weak property. The property simply requires that the social utility be monotone in individual preference orderings (see Moulin, 1988, p. 33).

Definition 9. Let $\beta^o \in \mathcal{B}$ be given. A social utility function u satisfies *unanimity* if $\forall \succcurlyeq \in \mathit{supp}(\beta^o)$,

$$[\forall i \in N, z \succcurlyeq_i z';\ \exists j \in N \text{ such that } z \succ_j z'] \quad \Rightarrow \quad \big[u(z, \succcurlyeq) > u(z', \succcurlyeq)\big].$$

[6] Note that condition (ii) in Definition 1A does not imply that posterior beliefs are necessarily degenerate. For example, if $g(s) \in \varphi(\succcurlyeq) \cap \varphi(\succcurlyeq')$ for any $\succcurlyeq, \succcurlyeq' \in \mathit{supp}(\beta^o)$, then $\beta(\cdot, s)$ may be non-degenerate.

Next, we define an alternative property of social utility functions which is natural in a context in which either the planner is assumed to be benevolent towards agents' welfare or it is common knowledge that subsequent to any inefficient allocation of resources by the designer, the agents can costlessly renegotiate and arrive at an efficient allocation through a Pareto-improving trade. Hence, the designer's utility from assigning an inefficient allocation cannot exceed that from assigning an efficient one.

Definition 10. Suppose $\beta^o \in \mathcal{B}$ is given. A social utility function u satisfies *limited preference for Pareto-efficiency (PPE)* if $\forall \succcurlyeq \in supp(\beta^o)$,

$$\forall z \notin PE(\succcurlyeq),\ \exists z' \in PE(\succcurlyeq) \text{ such that } \forall i \in N,\ z' \succcurlyeq_i z \text{ and } u(z', \succcurlyeq) \geqslant u(z, \succcurlyeq).$$

The following results show that, provided the social utility function satisfies unanimity, then the negative result relating to implementability of individually rational and efficient social choice rules persists. The social choice rules must satisfy a mild requirement of containing some strictly individually rational allocations.

Theorem 8. *If*

(i) $\varphi \subseteq PE \cap IR$,
(ii) $\forall \succcurlyeq \in \mathcal{R}$, $int[IR(\succcurlyeq)] \neq \emptyset \Rightarrow \varphi(\succcurlyeq) \cap int[IR(\succcurlyeq)] \neq \emptyset$,
(iii) $\mathcal{B}$ *is unrestricted, and*
(iv) u *satisfies unanimity,*

then φ *cannot be weakly credibly implemented in* $(\mathcal{B}, u)$.

Proof. Consider $\beta^o : \mathcal{R} \to [0, 1]$ such that $\beta^o(\succcurlyeq) = \beta^o(\succcurlyeq') = 0.5$ and the $PE \cap IR$ correspondence defined for $\succcurlyeq$ and $\succcurlyeq'$ is as given in Fig. 4. The broken-line indifference curves represent the economy $\succcurlyeq$, whereas the unbroken-line curves represent $\succcurlyeq'$. For each economy, the indifference curves are derived by horizontal translations of the curves shown. Denote $PE \cap IR(\succcurlyeq)$ by AB and $PE \cap IR(\succcurlyeq')$ by BC. The example is for a two-person, two-good economy. However, it easily generalizes.

We begin by supposing that φ is weakly credibly implementable in $(\mathcal{B}, u)$. We shall arrive at a contradiction in two steps. Let $\Gamma = \langle S, g \rangle$ be a game form that weakly credibly implements φ.

(1) First observe that, by construction, the unanimity condition on u implies that for any $z \in A$, if $z \in g(s)$ then $z \in \widetilde{PE}(\beta^o) = DE \cup DO_2 \cup EO_1$. Suppose otherwise, i.e. $z \notin DE \cup DO_2 \cup EO_1$. In this case, for both economies $\succcurlyeq$ and $\succcurlyeq'$, there will exist $z' \in DE \cup DO_2 \cup EO_1$ such that for $i = 1, 2$, $z'_i \succcurlyeq z$ and $z'' \succcurlyeq z$ and for each $\succcurlyeq^* \in \{\succcurlyeq, \succcurlyeq'\}$, for at least one agent i, $z' \succcurlyeq^*_i z$. This follows from the construction of the indifference curves in Fig. 4. Recall that, for each economy, the indifference curves are horizontal translations of the curves shown. Thus, by monotonicity of u, $\tilde{u}(z', \beta) > \tilde{u}(z, \beta)$ for all β. This is in contradiction with the assumption that Γ weakly credibly implements φ.

(2) Next, we shall argue that if Γ weakly credibly implements φ, then it must be the case that $\varphi(\succcurlyeq) \subseteq \varphi(\succcurlyeq')$. This is clearly, not the case in our example, except if $\tilde{\varphi}(\beta^o) = \{B\}$. This is in contradiction with the assumption of the theorem that states that $\varphi(\succcurlyeq) \cap int[IR(\succcurlyeq)] \neq \emptyset$.

To show that $\varphi(\succcurlyeq) \subseteq \varphi(\succcurlyeq')$, choose $\sigma \in NE(\Gamma)$ such that $g(\sigma(\succcurlyeq)) = z \in \varphi(\succcurlyeq)$. By the definition of a Nash equilibrium, for all $i \in N$, $\{z' \in A\colon \exists s(\succcurlyeq) \in S_i$ such that $g(s(\succcurlyeq), \sigma_{-i}(\succcurlyeq)) = z'\} \subseteq L_i(z, \succcurlyeq)$.

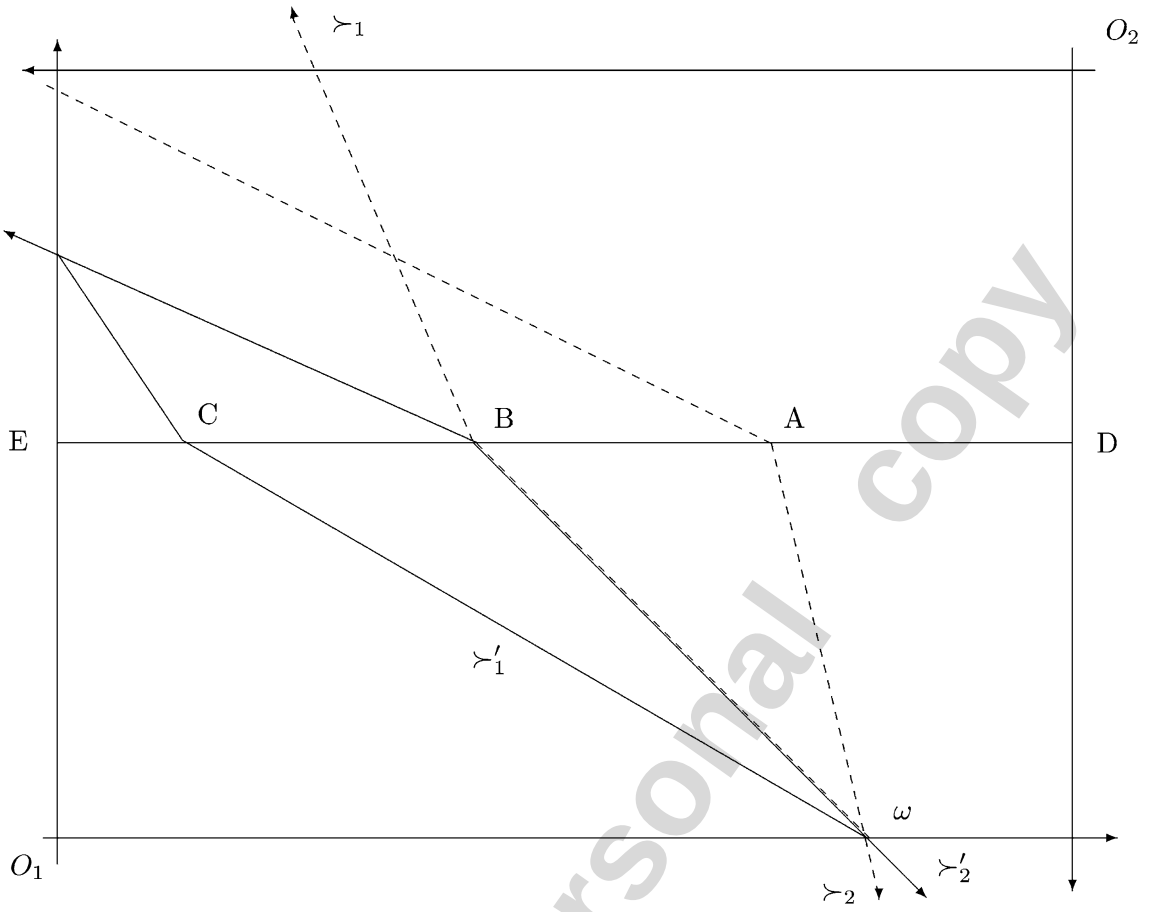

Fig. 4.

Consider the second economy $\succcurlyeq'$ and observe that, by construction, $\{L_i(z, \succcurlyeq) \cap \widetilde{PE}(\beta^o) \subseteq L_i(z, \succcurlyeq')$ for all $i \in N$. By step 1 above, $\{z' \in A\colon \exists s_i' \in S_i$ such that $g(s_i', \sigma_{-i}(\succcurlyeq)) = z'\} \subseteq \{L_i(z, \succcurlyeq) \cap \widetilde{PE}(\beta^o)\}$. Since $\{L_i(z, \succcurlyeq) \cap \widetilde{PE}(\beta^o)\} \subseteq L_i(z, \succcurlyeq')$, it must be the case that $g(\sigma(\succcurlyeq)) \in NE_A(\Gamma, \succcurlyeq')$. By definition of weakly credible implementability, $z = g(\sigma(\succcurlyeq)) \in \varphi(\succcurlyeq')$. □

The intuition underlying the example above is similar to that of the examples of the previous section. Once again, we have a planning problem in which the designer has two elements in the support of her prior beliefs about the true economy; this limits her ability to construct an outcome function with the appropriate incentives. Moreover since the individually rational and Pareto-efficient sets for the two economies are connected, regardless of the function u that is known to represent the designer's preferences, $\tilde{\varphi}(\beta^o)$ must be contained within this connected set, provided u is monotone. Given this observation the argument virtually mimics that of the previous section.

Next, suppose that the designer has a limited preference for efficiency. Under this assumption, the negative conclusions of the previous theorem can be obtained as a corollary of the previous result.

Theorem 9. *If*

(i) $\varphi \subseteq PE \cap IR$,
(ii) $\forall \succcurlyeq \in \mathcal{R}$, $int[IR(\succcurlyeq)] \neq \emptyset \Rightarrow \varphi(\succcurlyeq) \cap int[IR(\succcurlyeq)] \neq \emptyset$,

(iii) *$\mathcal{B}$ is unrestricted, and*
(iv) *u satisfies PPE,*

then φ cannot be weakly credibly implemented in $(\mathcal{B}, u)$.

Proof. Consider the example given in the proof of the previous theorem. The argument is identical to that given in the proof above, with the exception that the *PPE* property is used to prove step 1, instead of unanimity. □

The example given in this section is not pathological since it is designed to apply to every sub-correspondence of $PE \cap IR$. We used an example with certain special properties. But for any particular φ (e.g. the Walrasian correspondence) it can be seen that the non-implementability argument may be given using modifications of the arguments given in the proofs of Theorems 8 and 9.

5. Conclusions

This paper introduces the notion of credible implementation. We explicitly model the planner's preferences via the social choice correspondence and her beliefs about the true state of the economy, both of which are assumed to be common knowledge. Two notions of credibility are proposed. "Credible implementation" requires that, given a social choice rule, φ, for every prior in the domain of the planner's priors, there must exist a game form whose Nash equilibrium allocations coincide with the φ-optimal allocations for each economy in the support of the prior; moreover, the outcome function must choose only allocations that are φ-optimal for *some* economy in the support of the prior. A weaker definition of credible implementation is given in Section 4 by assuming that the social choice rule is derived from an underlying social utility function which is common knowledge. In the latter case, given prior beliefs, the outcome function of the implementing game form must choose allocations that are best-responses according to the known utility function and some posterior belief consistent with the prior.

Several other variations on the definitions are possible. Consider the following examples.

(i) The rule φ may be credibly implementable only for some priors but not for all priors as our definition requires; or
(ii) the planner's priors are not assumed common knowledge, in which case a single game form must credibly implement φ for every conceivable prior; or
(iii) given the prior β^o the choices of the planner might be restricted to be close to $\varphi(supp(\beta^o))$; or
(iv) the planner can only choose allocations in the convex hull of $\varphi(supp(\beta^o))$.

We did not pursue these variations here. Variation (i) would permit credible implementation of φ in some planning problems but not in others. This variation calls for an investigation on the set of admissible domains that is beyond the scope of this paper. In any case, the domains used in our negative results do not seem far fetched. Variation (ii), on the other hand, is too strong. We have shown that the results on credible implementability (as defined in this paper) are negative. Under this stronger notion, our negative results are likely to persist. Variation (iii) is similar to the idea of virtual implementation (Matsushima, 1988; Abreu and Sen, 1991). It is likely to produce more positive results than the other two. The last variation is motivated by the

fact that, if the planner is uncertain about the true economy she may choose allocations that are a convex combination of ex post optimal allocations. This notion gives the planer the maximum set of allocations she can choose from if she is restricted to ex ante optimal allocations. It is also likely to produce positive results since it enlarges the set of allocations that can be used to construct the mechanism.

Using the definitions proposed in the paper, we establish the following results. The positive findings of Nash implementation theory are severely affected once a credibility restriction is imposed. Recall that the Pareto correspondence and several sub-correspondences of the Pareto-efficiency and individual rationality correspondence, such as the (constrained) Walrasian correspondence, are Nash implementable. We show that

(a) the non-dictatorial Pareto correspondence can be credibly implemented,
(b) there exists no Pareto-efficient and individually rational social choice rule that can be implemented in a credible manner, and
(c) there exists no Pareto-efficient and envy-free social choice rule that can be implemented in a credible manner.

We also establish necessary conditions for credible implementability and derive conditions on the designer's beliefs that are sufficient for credible implementability. An implication of these conditions is that the less information a designer has, the easier it is to achieve her objective. The intuition behind this is clear from our conditions: a designer with relatively little information has a wider support for her priors—which in turn gives her more room to design a mechanism, since the outcome function can map to a larger subset of allocations.

Finally, the paper explores the robustness of the negative conclusions by weakening the credibility requirement. Such a weaker definition would require additional information about the designer's preferences, in excess of what is normally available in a standard social choice problem. We find that there exists no Pareto-efficient and individually rational social choice rule that can be implemented in a weakly credible manner under natural assumptions about the designer's preferences. The negative findings do not disappear even if the designer's preferences have some "continuity" properties.

We have formulated the idea of credible implementation using Nash equilibrium as the solution concept, primarily because the Nash model has been the starting point of a large number of papers in implementation theory. The concerns raised here also arise in other contexts such as implementation using solution concepts that refine Nash equilibria and in asymmetric information environments. In the case of the former, the results of Moore and Repullo (1988), Palfrey and Srivastava (1991) and Palfrey et al. (1994) have shown that "almost" all social choice correspondences are implementable using alternative refinements. On the other hand, by weakening the requirement of exactness of implementation, Matsushima (1988) and Abreu and Sen (1991) have demonstrated similarly startling results. A natural line of inquiry is: how are these strong results limited by a restriction of credibility? We leave this question for future work. In asymmetric information domains, the results have been largely negative (Palfrey and Srivastava, 1987; Chakravorti, 1992, 1993) to begin with. However, given the satisfaction of an incentive compatibility requirement, positive results emerge. These are, in fact, made stronger by the use of refinements (Palfrey and Srivastava, 1991). Here again, we must study the impact of credibility.

We close with the following observations:

(i) The credibility issue is especially severe for implementation theorems that use assumptions such as the existence of a universally worst element (e.g. see Palfrey et al., 1994 on bounded

implementation or the papers on Bayesian implementation). It is hard to imagine that a designer (whose preferences are given by a reasonable social choice rule) can rationalize an outcome in which she destroys all the goods in an economy. Even though in equilibrium, such a threat is never carried out, it is an incredible one and will make the rationale leading to an equilibrium unravel.

(ii) Our concern for out-of-equilibrium moves has been voiced in the past in the context of well-behavedness of the outcome function. The focus has been on continuity (Postlewaite and Wettstein, 1989) and feasibility (Hurwicz et al., 1995). The issue of credibility has not been addressed, though it may be tied into the concern for continuity and feasibility.

(iii) Note that a revelation principle can be stated where we replace "implementability" with "credible implementability." The principle essentially states that any social choice that can be implemented using an arbitrary game form can also be truthfully implemented using a direct revelation mechanism. But a direct mechanism, appropriately defined, is always credible provided its outcome function is a selection from the correspondence φ. But that is exactly the manner in which a direct mechanism is constructed to prove the principle.

(iv) In classical social choice theory, a condition that is often assumed is that of an "unrestricted domain" (of agents' preferences). Of course, under such an assumption, the counterexamples presented here would not arise. The role of this assumption in the social choice literature is to demonstrate that particular results hold in a wide universe of problems. Our purpose is quite different. We argue that when one considers different subsets of such a wide universe, the results are crucially affected. The principal characteristic of such subsets is that the designer has some limited knowledge of the possible class of preferences, which enables her to eliminate some parts of the feasible set from being possible φ-optima. If the principal has *some* knowledge about the sub-class of preferences, which is available through experience or relatively costless surveys, then the concerns raised in this paper may be relevant.

Acknowledgments

The authors would like to thank D. Abreu, P. Amorós, C. Beviá, S. Srivastava, two anonymous referees and an associate editor for their detailed comments and suggestions. They are also grateful, for the many helpful discussions, to participants of the Mid-West Mathematical Economics Conference at Indianapolis, the International Conference on Game Theory at Stony Brook, the Econometric Society meeting in Boston, and seminar audiences at the following universities: Alicante, Barcelona, Columbia, California at Davis, Harvard, the Hebrew University, Illinois at Urbana-Champaign and Texas at Austin.

References

Abreu, D., Sen, A., 1991. Virtual implementation in Nash equilibrium. Econometrica 59, 997–1021.

Aghion, P., Dewatripont, M., Rey, P., 1994. Renegotiation design with unverifiable information. Econometrica 62, 257–282.

Baliga, S., Sjostrom, T., 1999. Interactive implementation. Games Econ. Behav. 27, 38–63.

Baliga, S., Corchón, L., Sjostrom, T., 1997. The theory of implementation when the planner is a player. J. Econ. Theory 77, 15–33.

Chakravorti, B., 1992. Efficiency and mechanisms with no regret. Int. Econ. Rev. 33, 45–60.

Chakravorti, B., 1993. Sequential rationality, implementation and pre-play communication. J. Math. Econ. 22, 265–294.

Gevers, L., 1986. Walrasian social choice: Some simple axiomatic approaches. In: Heller, W., Starr, R., Starrett, D. (Eds.), Social Choice and Public Decision Making: Essays in Honor of Kenneth Arrow, vol. 1. Cambridge Univ. Press, Cambridge, UK.

Hurwicz, L., 1979. Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. Rev. Econ. Stud. 46, 217–225.

Hurwicz, L., Maskin, E., Postlewaite, A., 1995. Feasible Nash implementation of social choice rules when the designer does not know endowments or production sets. In: Ledyard, J.O. (Ed.), The Economics of Information Decentralization: Complexity, Efficiency and Stability. Kluwer, Amsterdam, pp. 367–433.

Jackson, M., Palfrey, T., 2001. Voluntary implementation. J. Econ. Theory 98, 1–25.

Maskin, E., 1985. The theory of implementation in Nash equilibrium. In: Hurwicz, L., Schmeidler, D., Sonnenschein, H. (Eds.), Social Goals and Social Organizations: Essays in Memory of Elisha Pazner. Cambridge Univ. Press, Cambridge.

Maskin, E., 1999. Nash equilibrium and welfare optimality. Rev. Econ. Stud. 66, 23–38.

Maskin, E., Moore, J., 1999. Implementation with renegotiation. Rev. Econ. Stud. 66, 83–114.

Matsushima, K., 1988. A new approach to the implementation problem. J. Econ. Theory 45, 128–144.

McKelvey, R., 1989. Game forms for Nash implementation of social choice rules. Soc. Choice Welfare 6, 139–156.

Moore, J., Repullo, R., 1988. Subgame perfect implementation. Econometrica 56, 1191–1220.

Moulin, H., 1988. Axioms of Cooperative Decision Making. Econometric Society Monographs. Cambridge Univ. Press, Cambridge, UK.

Palfrey, T., Srivastava, S., 1987. On Bayesian implementable allocations. Rev. Econ. Stud. LIV, 193–208.

Palfrey, T., Srivastava, S., 1991. Nash implementation using undominated strategies. Econometrica 59, 479–501.

Palfrey, T., Srivastava, S., Jackson, M., 1994. Undominated Nash implementation in bounded mechanisms. Games Econ. Behav. 6, 474–501.

Postlewaite, A., Wettstein, D., 1989. Implementing constrained Walrasian equilibria continuously. Rev. Econ. Stud. 56, 603–611.

Ray, D., Ueda, K., 1996. Egalitarianism and incentives. J. Econ.Theory 71, 324–348.

Rubinstein, A., Wolinsky, A., 1992. Renegotiation-proof implementation and time preferences. Amer. Econ. Rev. 82, 600–614.

Thomson, W., 1984. The manipulability of resource allocation mechanisms. Rev. Econ. Stud. 51, 447–460.

Thomson, W., 1987. The vulnerability to manipulative behavior of resource allocation mechanisms designed to select equitable and efficient outcomes. In: Groves, T., Radner, R., Reiter, S. (Eds.), Information, Incentives and Economic Mechanisms: Essays in Honor of Leonid Hurwicz. Univ. of Minnesota Press, Minneapolis.