ROBUST IMPLEMENTATION IN GENERAL MECHANISMS

By

Dirk Bergemann and Stephen Morris

June 2008
Revised January 2009

COWLES FOUNDATION DISCUSSION PAPER NO. 1666R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Robust Implementation in General Mechanisms*

Dirk Bergemann† Stephen Morris‡

January 2010

Abstract

A social choice function is *robustly implemented* if every equilibrium on every type space achieves outcomes consistent with it. We identify a *robust monotonicity* condition that is necessary and (with mild extra assumptions) sufficient for robust implementation.

Robust monotonicity is strictly stronger than both Maskin monotonicity (necessary and almost sufficient for complete information implementation) and ex post monotonicity (necessary and almost sufficient for ex post implementation). It is equivalent to Bayesian monotonicity on all type spaces.

KEYWORDS: Mechanism Design, Implementation, Robustness, Common Knowledge, Interim Equilibrium, Dominant Strategies.

JEL CLASSIFICATION: C79, D82

*This research is supported by NSF Grants #CNS-0428422 and #SES-0518929. We thank Matt Jackson, the co-editor, Andy Postlewaite and two anonymous referees for helpful comments. This paper supersedes and incorporates results reported earlier in Bergemann and Morris (2005a). Through our joint authorship of Bergemann, Morris and Tercieux (2010) and detailed comments on this draft, Olivier Tercieux has greatly improved this version, including suggesting a strengthening of Theorem 1 and the treatment of responsive social choice functions in Section 6.1.

†Department of Economics, Yale University, New Haven, CT 06511, dirk.bergemann@yale.edu

‡Department of Economics, Princeton University, Princeton, NJ 08544, smorris@princeton.edu

1 Introduction

The objective of mechanism design is to construct mechanisms (or game forms) such that privately informed agents have an incentive to reveal their information to a principal who seeks to realize a social choice function. The revelation principle establishes that if any mechanism can induce the agents to report their information, then the agents will also have an incentive to report truthfully in the direct mechanism. Given the beliefs of the agents, the truthtelling constraints then reduce in the direct mechanism to the Bayesian incentive compatibility conditions.

There are two important limitations of Bayesian incentive compatibility analysis. First, the analysis typically assumes a commonly known common prior over the agents' types. This assumption may be too stringent in practise. In the spirit of the "Wilson doctrine" (Wilson (1987)), we would like implementation results that are *robust* to different assumptions about what agents do or do not know about other agents' types. Second, the revelation principle only establishes that the direct mechanism has *an* equilibrium that achieves the social choice function. In general, there may be other interim (or Bayesian) equilibria that deliver undesirable outcomes.[1] We would like to achieve *full* implementation, i.e., show the existence of a mechanism all of whose interim equilibria deliver the social choice function. We studied the first "robustness" problem in an earlier work, Bergemann and Morris (2005). The second "full implementation" problem has been the subject of a large literature. In the incomplete information context, key full implementation references are Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989) and Jackson (1991). In this paper, we study "robust implementation" where we require robustness and full implementation simultaneously.

The notion of robust implementation requires that a social choice function f can be interim implemented for all type spaces $\mathcal{T}$. As we look for necessary and sufficient conditions for robust implementation, conceptually there are (at least) two approaches to obtain the conditions. One approach would be to simply look at the interim implementation conditions for every possible type space $\mathcal{T}$ and then try to characterize the intersection or union of these conditions for all type spaces. This is the approach we initially pursued, and it works in a brute force kind of way. In Section 6.3, we review what happens under this approach. But we focus our analysis on a second, more elegant, approach. We first establish an equivalence between robust and *rationalizable implementation* and then derive the conditions for robust implementation as an implication of rationalizable implementation. The advantage of the second approach is that after establishing the equivalence, we do not need to argue in terms of large type spaces, but rather derive the results from a novel

[1] We typically refer to "interim" equilibria rather than "Bayesian" equilibria in light of the fact that the type space may not necessarily have a common prior.

argument using the iterative deletion procedure associated with rationalizability. For *rationalizable implementation*, we fix a mechanism and iteratively delete messages for each payoff type that are strictly dominated by another message for each payoff type profile and message profile that has survived the procedure.[2] The equivalence between robust and rationalizable implementation illustrates a general point well-known from the literature on epistemic foundations of game theory (e.g., Brandenburger and Dekel (1987), Battigalli and Siniscalchi (2003)): equilibrium solution concepts only have bite if we make strong assumptions about type spaces, i.e., we assume small type spaces where the common prior assumption holds.

We exploit the equivalence between robust and rationalizable implementation to obtain necessary and sufficient conditions for robust implementation in general environments. Our necessity argument is conceptually novel, exploiting the iterative characterization. The necessary conditions for robust implementation are ex post incentive compatibility of the social choice function and a condition - *robust monotonicity* - that is equivalent to requiring Bayesian monotonicity on every type space. This condition is strong and implies Maskin monotonicity - necessary and almost sufficient for complete information implementation - but is strictly weaker than dominant strategy implementation.

The sufficiency argument requires only a modest strengthening of the necessary condition by guaranteeing that the preference profile of each agent satisfies a (conditional) no total indifference property. Under this no total indifference property, we show that the necessary conditions are also sufficient for robust implementation. The sufficient conditions guarantee robust implementation in pure, but also in mixed strategies. Our robust analysis thus removes the frequent gap between pure and mixed strategy implementation in the literature.

In this paper, we follow the classic implementation literature in allowing for arbitrary mechanisms, including modulo and integer games. By allowing for these mechanisms, we are able to make tight connections with the existing implementation literature. Allowing for these badly behaved mechanisms does complicate our analysis: for example, we must allow for transfinite iterated deletion of best responses in our definition of rationalizable implementation. Given the complications arising from infinite mechanisms, we report new necessary conditions for robust implementation in the context of finite mechanisms. We also report how our earlier research can be used to show that these necessary conditions are sufficient conditions for finite mechanisms either in well-behaved, but restricted, environments (Bergemann and Morris (2009a)) or under a virtual rather than exact

[2] We use rationalizability (and rationalizable implementation) in this paper to refer to the "robust" version of the solution concept where payoff types are not known but any beliefs about others' payoff types are considered. We will later discuss the relation to implementation in rationalizable strategies in complete information settings (Bergemann, Morris, and Tercieux (2010)) and interim versions of rationalizability (Dekel, Fudenberg, and Morris (2007)).

implementation requirement (Bergemann and Morris (2009b)).

Our results extend the classic literature on Bayesian implementation due to Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989) and Jackson (1991). We focus in this paper on an indirect approach to extending these results. We first note the equivalence between robust implementation and rationalizable implementation. We then exploit the equivalence to report a direct argument showing that robust monotonicity is a necessary and almost sufficient condition for rationalizable implementation. But in the light of the classic literature, we know that a necessary and almost sufficient condition for robust implementation must be Bayesian monotonicity on all type spaces. We confirm and clarify our results by directly checking that robust monotonicity is equivalent to Bayesian monotonicity on all type spaces. Figure 1 gives a stylized account of the connection between these alternative approaches.

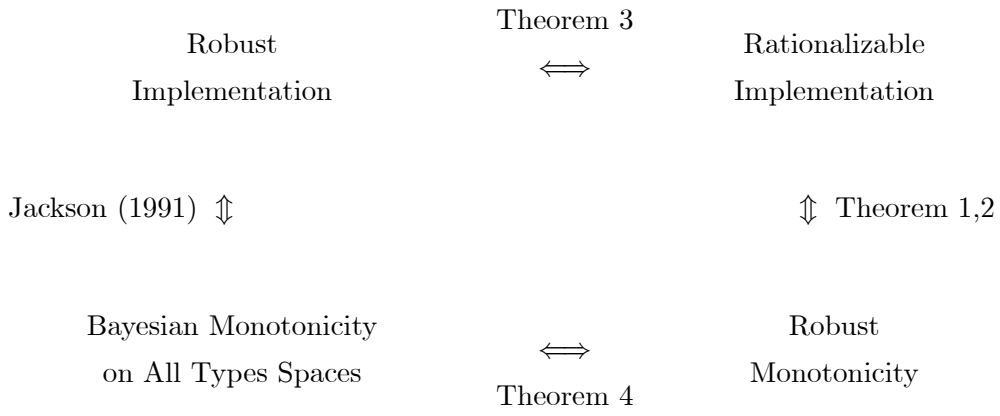

Figure 1: Relationship between Bayesian and Robust Implementation / Monotonicity

In the implementation literature, it is a standard practice to obtain the sufficiency results with augmented mechanisms. By augmenting the direct mechanism with additional messages, the designer may elicit additional information about undesirable equilibrium play by the agents. Yet, in many applied economic settings, single crossing or supermodular preference assumptions allow direct implementation. In a companion paper, Bergemann and Morris (2009a), we provide necessary and sufficient conditions for robust implementation in the *direct mechanism*. The main results there apply to environments where each agent's type profile can be aggregated into a one dimensional sufficient statistic and where preferences are single crossing with respect to that statistic. These restrictions incorporate many economic models with interdependence in the literature. We show that besides an incentive compatibility condition, in this case the strict ex post incentive compatibility condition, a *contraction* property which requires that there is *not too much* interdependence in agents' types, together present necessary and sufficient conditions for robust implementation in

the direct mechanism.

The robust monotonicity condition is stronger than both the Maskin and the Bayesian monotonicity conditions. In the context of robust implementation, we can ask whether a relaxation from the exact to the virtual implementation condition may lead to more permissive results. In Bergemann and Morris (2009b) we characterize the necessary and sufficient conditions for *robust virtual* implementation. There we show that a social choice function can be *robustly virtually implemented* if and only if the social choice function is ex post incentive compatible and robust measurable. We establish here that robust measurability remains a necessary condition for robust (exact) implementation, but it is not sufficient anymore.

The results in this paper concern full implementation. An earlier paper of ours, Bergemann and Morris (2005), addresses the analogous questions of robustness to rich type spaces, but looking at the question of partial implementation, i.e., does there exist a mechanism such that *some* equilibrium implements the social choice function. We showed that ex post (partial) implementation of the social choice function is a necessary and sufficient condition for partial implementation on all type spaces.[3] This paper establishes that an analogous result does *not* hold for full implementation. In a related paper, Bergemann and Morris (2008a), we therefore investigate the notion of ex post implementation. The necessary and sufficient conditions there straddle the implementation conditions for Nash and Bayesian-Nash respectively, as an ex post equilibrium is a Nash equilibrium at every incomplete information (Bayesian) type profile. However in contrast to the iterative argument pursued here, the basic reasoning in Bergemann and Morris (2008a) invokes more traditional equilibrium arguments. By comparing the conditions for ex post and robust implementation, it becomes apparent that robust implementation typically imposes additional constraints on the allocation problem. In Bergemann and Morris (2008a), we showed that in single crossing environments, the same single crossing conditions which guarantee incentive compatibility also guarantee full ex post implementation. In contrast, in the aggregation environment discussed above, we show that robust implementation imposes a strict bound on the interdependence of the preferences, which is not required by the truthtelling conditions. A contraction mapping behind the iterative argument directly points to the source of the restriction of the interaction term.

Our results provide a characterization of rationalizable implementation in incomplete information environments. In a recent work, Bergemann, Morris, and Tercieux (2010), we have adapted the arguments to characterize rationalizable implementation in complete information environments.

The remainder of the paper is organized as follows. Section 2 describes the formal environment and solution concepts. Section 3 gives necessary and sufficient conditions for rationalizable imple-

[3] This result does not extend to social choice correspondences.

mentation. Section 4 establishes an (almost) equivalence relation between rationalizable and robust implementation and then reports the necessary and sufficient conditions for robust implementation. Section 5 establishes necessary conditions for robust implementation in finite mechanisms. Section 6 discusses extensions and variations of our implementation results, examining the role of lotteries and pure strategies and the relationship with Nash equilibrium and ex post equilibrium implementation. The Appendix contains some additional examples.

2 Setup

2.1 The Payoff Environment

We consider a finite set of agents, $1, 2, ..., I$. Agent i's *payoff type* is $\theta_i \in \Theta_i$. We write $\theta \in \Theta = \Theta_1 \times \cdots \times \Theta_I$. There is a set of outcomes Z. We assume that each Θ_i and Z are countable.[4] Each individual has a von Neumann-Morgenstern utility function $u_i : Z \times \Theta \to \mathbb{R}$. Thus we are in the world of interdependent types, where an agent's utility depends on other agents' payoff types. We allow for lotteries over deterministic outcomes.[5] Let $Y \triangleq \Delta(Z)$ and extend u_i to the domain $Y \times \Theta$ in the usual way:

$$u_i(y, \theta) \triangleq \sum_{z \in Z} y(z)\, u_i(z, \theta).$$

A social choice function is a mapping $f : \Theta \to Y$. If the true payoff type profile is θ, the planner would like the outcome to be $f(\theta)$. In this paper, we restrict our analysis to the implementation of a social choice function rather than a social choice correspondence or social choice set.[6]

2.2 Type Spaces

We are interested in analyzing behavior in a variety of type spaces, many of them with a richer set of types than payoff types. For this purpose, we shall refer to agent i's *type* as $t_i \in T_i$, where T_i is a countable set. A type of agent i must include a description of his payoff type. Thus there is a function $\widehat{\theta}_i : T_i \to \Theta_i$ with $\widehat{\theta}_i(t_i)$ being agent i's payoff type when his type is t_i. A type of agent i must also include a description of his beliefs about the types of the other agents; thus there is a function $\widehat{\pi}_i : T_i \to \Delta(T_{-i})$ with $\widehat{\pi}_i(t_i)$ being agent i's *belief type* when his type is t_i. Thus

[4] The countable types restriction clarifies the relation to the existing literature. We briefly discuss what happens if we allow for uncountable payoff types, types and pure outcomes in Section 6.5.

[5] The role of the lottery assumption and what happens when we drop it are discussed in Section 6.3.

[6] One reason why the extension to social choice correspondences is not straightforward is that, with social choice correspondences, the incentive compatibility conditions that arise from requiring partial implementation are typically weaker than ex post incentive compatibility, as shown by examples in Bergemann and Morris (2005).

$\widehat{\pi}_i(t_i)[t_{-i}]$ is the probability that type t_i of agent i assigns to other agents having types t_{-i}. A *type space* is a collection:

$$\mathcal{T} = \left(T_i, \widehat{\theta}_i, \widehat{\pi}_i\right)_{i=1}^{I}.$$

2.3 Mechanisms

A planner must choose a *game form* or *mechanism* for the agents to play in order to determine the social outcome. Let M_i be the countably infinite set of messages available to agent i. We denote the generic message by $m_i \in M_i$ and let $m \in M = M_1 \times \cdots \times M_I$. Let $g(m)$ be the distribution over outcomes if action profile m is chosen. Thus a mechanism is a collection $\mathcal{M} = (M_1, ..., M_I, g(\cdot))$, where $g : M \to Y$.

2.4 Solution Concepts

Now holding fixed the payoff environment, we can combine a type space $\mathcal{T}$ with a mechanism $\mathcal{M}$ to get an incomplete information game $(\mathcal{T}, \mathcal{M})$. The payoff of agent i if message profile m is chosen and type profile t is realized is then given by

$$u_i\left(g(m), \widehat{\theta}(t)\right).$$

A pure strategy for agent i in the incomplete information game $(\mathcal{T}, \mathcal{M})$ is given by

$$s_i : T_i \to M_i.$$

A (behavioral) strategy is given by

$$\sigma_i : T_i \to \Delta(M_i).$$

The objective of this paper is to obtain implementation results for interim, or Bayesian Nash, equilibria on all possible types spaces. The notion of interim equilibrium for a given type space $\mathcal{T}$ is defined in the usual way.

Definition 1 (Interim equilibrium)
A strategy profile $\sigma = (\sigma_1, ..., \sigma_I)$ is an interim equilibrium of the game $(\mathcal{T}, \mathcal{M})$ if, for all i, t_i and m_i with $\sigma_i(m_i|t_i) > 0$,

$$\begin{aligned} & \sum_{t_{-i} \in T_{-i}} \sum_{m_{-i} \in M_{-i}} \left(\prod_{j \neq i} \sigma_j(m_j|t_j) \right) u_i\left(g(m_i, m_{-i}), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_i)[t_{-i}] \\ \geq & \sum_{t_{-i} \in T_{-i}} \sum_{m_{-i} \in M_{-i}} \left(\prod_{j \neq i} \sigma_j(m_j|t_j) \right) u_i\left(g(m_i', m_{-i}), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_i)[t_{-i}], \quad \forall m_i'. \end{aligned}$$

Requiring "robust" implementation, i.e., for "all type spaces", will push the solution concept in the direction of rationalizability. Consequently, we define a message correspondence profile $S = (S_1, ..., S_I)$, where each

$$S_i : \Theta_i \to 2^{M_i}, \tag{1}$$

and we write $\mathcal{S}$ for the collection of message correspondence profiles. The collection $\mathcal{S}$ is a lattice with the natural ordering of set inclusion: $S \leq S'$ if $S_i(\theta_i) \subseteq S'_i(\theta_i)$ for all i and θ_i. The largest element is $\overline{S} = \left(\overline{S}_1, ..., \overline{S}_I\right)$, where $\overline{S}_i(\theta_i) = M_i$ for all i and θ_i. The smallest element is $\underline{S} = (\underline{S}_1, ..., \underline{S}_I)$, where $\underline{S}_i(\theta_i) = \varnothing$ for all i and θ_i.

We define an operator $b = (b_1, ..., b_I)$ to iteratively eliminate never best responses. To this end, we denote the belief of agent i over message and payoff type profiles of the remaining agents by

$$\lambda_i \in \Delta(M_{-i} \times \Theta_{-i}).$$

The operator $b : \mathcal{S} \to \mathcal{S}$ is now defined as:

$$b_i(S)[\theta_i] \triangleq \left\{ m_i \in M_i \,\middle|\, \exists \lambda_i \text{ s.th.:} \begin{array}{ll} (1) & \lambda_i(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j(\theta_j) \text{ for all } j \neq i; \\ (2) & \sum\limits_{m_{-i}, \theta_{-i}} \lambda_i(m_{-i}, \theta_{-i}) u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i})) \geq \\ & \sum\limits_{m_{-i}, \theta_{-i}} \lambda_i(m_{-i}, \theta_{-i}) u_i(g(m'_i, m_{-i}), (\theta_i, \theta_{-i})), \ \forall m'_i \in M_i; \end{array} \right\}.$$

We observe that b is increasing by definition: i.e., $S \leq S' \Rightarrow b(S) \leq b(S')$. By Tarski's fixed point theorem, there is a largest fixed point of b, which we label $S^{\mathcal{M}}$. Thus (i) $b\left(S^{\mathcal{M}}\right) = S^{\mathcal{M}}$ and (ii) $b(S) = S \Rightarrow S \leq S^{\mathcal{M}}$. We can also construct the fixed point $S^{\mathcal{M}}$ by starting with $\overline{S}$ - the largest element of the lattice - and iteratively applying the operator b. If the message sets and types are finite, we have

$$S_i^{\mathcal{M}}(\theta_i) \triangleq \bigcap_{n \geq 1} b_i\left(b^n\left(\overline{S}\right)\right)[\theta_i].$$

But because the mechanism $\mathcal{M}$ may be infinite, transfinite induction may be necessary to reach the fixed point.[7] It is useful to define

$$S_i^{\mathcal{M},k}(\theta_i) \triangleq b_i\left(b^{k-1}\left(\overline{S}\right)\right)[\theta_i],$$

again using transfinite induction if necessary. Thus $S_i^{\mathcal{M}}(\theta_i)$ are the set of messages surviving (transfinite) iterated deletion of never best responses; equivalently, $S_i^{\mathcal{M}}(\theta_i)$ is the set of messages that a player with payoff type θ_i might send consistent with common certainty of rationality, but

[7] Lipman (1994) contains a formal description of the transfinite induction required (for the case of complete information, but nothing important changes with incomplete information). As he notes "we remove strategies which are never a best reply, taking limits where needed".

no restrictions on higher order beliefs about others' types. We refer to $S_i^{\mathcal{M}}(\theta_i)$ as the *rationalizable* messages of type θ_i of agent i in mechanism $\mathcal{M}$.

If message sets are finite (or compact), a well known duality argument implies that never best responses are equivalent to strictly dominated actions. However, the equivalence does not hold with infinite (non-compact) message sets.[8] In a compact message analysis, Chung and Ely (2001) consider a version of this solution concept in an incomplete information mechanism design context with dominated (not strictly dominated) messages deleted at each round. The solution concept defined through the iterative application of the operator b is weaker than the notion of interim rationalizability for a given type space $\mathcal{T}$, as defined by Battigalli and Siniscalchi (2003) and Dekel, Fudenberg, and Morris (2007). Under b, every agent i is allowed to hold arbitrary beliefs about Θ_{-i} and is not restricted to a particular posterior distribution over Θ_{-i}. On the other hand, if the type space $\mathcal{T}$ were the universal type space, then $S_i^{\mathcal{M}}(\theta_i)$ would be equal to the union of all interim rationalizable actions of agent i over all types $t_i \in T_i$ whose payoff type profile coincides with θ_i, or $\widehat{\theta}_i(t_i) = \theta_i$.

2.5 Implementation

We now define the notions of interim, robust and rationalizable implementation.

Definition 2 (Interim Implementation)
Social choice function f is interim implemented on type space $\mathcal{T}$ by mechanism $\mathcal{M}$ if the game $(\mathcal{T}, \mathcal{M})$ has an equilibrium and every equilibrium σ of the game $(\mathcal{T}, \mathcal{M})$ satisfies

$$\sigma(m|t) > 0 \Rightarrow g(m) = f\left(\widehat{\theta}(t)\right).$$

A tradition in the implementation literature commonly restricts attention to pure strategy equilibria, but we allow mixed strategy equilibria.

Definition 3 (Robust Implementation)
Social choice function f is robustly implemented by mechanism $\mathcal{M}$ if, for every $\mathcal{T}$, f is interim implemented on type space $\mathcal{T}$ by mechanism $\mathcal{M}$. Social choice function f is robustly implementable if there exists a mechanism $\mathcal{M}$ such that f is robustly implemented by mechanism $\mathcal{M}$.

[8] The following example, suggested to us by Andrew Postlewaite, illustrates the non-equivalence. Players 1 and 2 each choose a non-negative integer, k_1 and k_2 respectively. The payoff to player 1 from $k_1 = 0$ is 1. The payoff to player 1 from action $k_1 \geq 1$ is 2 if $k_1 > k_2$, 0 otherwise. For any belief that player 1 has about 2's actions, there is a (sufficiently high) action from player 1 that gives him a payoff greater than 1. Thus action 0 is never a best response for player 1. However, for any mixed strategy of player 1, there is a (sufficiently high) action of player 2 such that action 0 is a better response for player 1 than the mixed strategy. Thus action 0 is not strictly dominated.

The notion of robust implementation requires that we can find a mechanism $\mathcal{M}$ which implements f for every type space $\mathcal{T}$. A weaker requirement would be to ask that for every type space $\mathcal{T}$ there exists a, possibly different, mechanism $\mathcal{M}$ such that f is implemented. This weaker notion would still lead to the same necessary condition as the stronger implementation version we pursue here, and we believe that it would not lead to a change in the sufficiency conditions either.

Definition 4 (Rationalizable Implementation)
Social choice function f is rationalizably implementable by mechanism $\mathcal{M}$ if

1. $m \in S^{\mathcal{M}}(\theta) \Rightarrow g(m) = f(\theta)$; *and*

2. *For each i and $\psi_i \in \Delta(\Theta_{-i})$, there exists $\lambda_i \in \Delta(M_{-i} \times \Theta_{-i})$ such that:*

$$\begin{aligned}
&(a)\ \lambda_i(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j^{\mathcal{M}}(\theta_j) \quad \text{for all } j \neq i, \\
&(b)\ \sum_{m_{-i}} \lambda_i(m_{-i}, \theta_{-i}) = \psi_i(\theta_{-i}) \quad \text{for all } \theta_{-i} \in \Theta_{-i}, \\
&(c)\ \arg\max_{m_i} \sum_{m_{-i}, \theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i})) \neq \varnothing \quad \text{for all } \theta_i \in \Theta_i.
\end{aligned}$$

Social choice function f is rationalizably implementable if there exists a mechanism $\mathcal{M}$ such that f is rationalizably implementable by $\mathcal{M}$.

Part (1) of the definition requires that every rationalizable message profile leads to an outcome consistent with the social choice function. Part (2) requires that rationalizable messages exist. It is automatically satisfied if $\mathcal{M}$ is finite. But existence is non-trivial if infinite mechanisms are allowed and thus best responses may not exist to all conjectures. The existence requirement is strong: we require that for each belief that agent i may have over the payoff types of the other agents, there exists a belief over the rationalizable messages that other agents might send such that agent i has a best response, whatever his payoff type. Following Theorem 1, we will report why our proof will not go through, in general, under weaker versions of this requirement; but in Section 6.1, we report how the best response property could be weakened in the special case where the social choice function is *responsive*, so that distinct payoff types of an agent always lead to distinct outcomes under the social choice function, for at least one payoff type profile of the other agents.

Note that if a best response exists for agent i to a conjecture over rationalizable messages, as in part (c), then by part (1), every rationalizable action of agent i will be a best response. Thus we could strengthen property (c) in the definition of rationalizable implementation to

$$\text{(c')}\ \arg\max_{m_i} \sum_{m_{-i}, \theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i})) = S_i^{\mathcal{M}}(\theta_i) \neq \varnothing \text{ for all } \theta_i \in \Theta_i,$$

without strengthening the definition.

3 Rationalizable Implementation

3.1 Necessity

The following ex post incentive compatibility condition is a necessary condition for robust truthful (or partial) implementation as established in Bergemann and Morris (2005).

Definition 5 (EPIC)
Social choice function f satisfies ex post incentive compatibility (EPIC) if

$$u_i\left(f\left(\theta_i,\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right)\geq u_i\left(f\left(\theta_i',\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right),$$

for all i, θ_i, θ_i' and θ_{-i}.

But a strengthening of this condition will be necessary for robust full implementation.

Definition 6 (Semi-Strict EPIC)
Social choice function f satisfies semi-strict ex post incentive compatibility (semi-strict EPIC) if, for each i, θ_i, θ_i' :

$$u_i\left(f\left(\theta_i,\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right)>u_i\left(f\left(\theta_i',\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right),\quad\forall\theta_{-i},$$

if there exists $\theta_{-i}'\in\Theta_{-i}$ such that $f\left(\theta_i,\theta_{-i}'\right)\neq f\left(\theta_i',\theta_{-i}'\right)$.

Next we present the monotonicity conditions which are at the core of the robust implementation results. It is useful to first think about agents playing the direct mechanism, where each agent i reports his payoff type. An agent i may or may not report truthfully. A *deception* is a set-valued profile $\beta=(\beta_1,...,\beta_I)$, where

$$\beta_i:\Theta_i\rightarrow 2^{\Theta_i}/\varnothing,$$

with $\theta_i\in\beta_i\left(\theta_i\right)$ for all i and all θ_i. A deception of agent i with payoff type θ_i is a set of possible reports by agent i. Thus a deception of payoff type θ_i includes, but is not restricted to, θ_i itself.

Definition 7 (Acceptable / Unacceptable Deception)
A deception β is acceptable if $\theta'\in\beta\left(\theta\right)\Rightarrow f\left(\theta'\right)=f\left(\theta\right)$. A deception is unacceptable if it is not acceptable.

In this language, the "truthtelling" deception, defined by $\beta_i^*\left(\theta_i\right)\triangleq\theta_i$ for all θ_i is an acceptable deception. Other deceptions of agent i may also be acceptable if the social choice function does not vary with respect to some subset of reports of agent i for all type profiles of the other agents. The

inverse mapping of a deception β_i represents the set of true type profiles θ_i which could lead to a report θ_i' and we write

$$\beta_i^{-1}\left(\theta_i'\right) \triangleq \left\{\theta_i \left| \theta_i' \in \beta_i\left(\theta_i\right)\right.\right\}.$$

A "robust monotonicity" condition is key to our main result. If a deception is "unacceptable," it must be possible to "refute" it. In the direct mechanism, where agents other than i report themselves to be types θ_{-i}, agent i can obtain outcomes $f\left(\theta_i', \theta_{-i}\right)$ for any θ_i'. But once we allow augmented mechanisms, we can offer agent i a larger set of lotteries if he reports deviant behavior of his opponents. We need to identify, for any given report θ_{-i} of agents other than i, the set of lotteries with the property that whatever agent i's actual type, he would never prefer such an allocation to what he would obtain under the social choice function if other agents were reporting truthfully. Thus:

$$Y_i\left(\theta_{-i}\right) \triangleq \left\{y \in Y \left| u_i\left(y, \left(\theta_i'', \theta_{-i}\right)\right) \leq u_i\left(f\left(\theta_i'', \theta_{-i}\right), \left(\theta_i', \theta_{-i}\right)\right) \text{ for all } \theta_i'' \in \Theta_i\right.\right\}. \tag{2}$$

Henceforth, we refer to the set $Y_i\left(\theta_{-i}\right)$ as the *reward set* (for agent i).

Suppose now that it was common knowledge that in the direct mechanism, type θ_i of agent i will send a report $\theta_i' \in \beta_i\left(\theta_i\right)$. But if β is not acceptable, we must find a type of some agent who is prepared to report that other agents are misreporting. But for the "whistle-blower" who is going to report that we are in a bad equilibrium, we cannot know what he believes about the types of the other agents, nor can we know what message he expects to hear except that it is a message consistent with the deception. We thus have to allow for all possible beliefs ψ_i of agent i over payoff types $\theta_{-i} \in \Theta_{-i}$ consistent with a report θ_{-i}' from a given deception profile β, or

$$\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right).$$

Finally, the reward that he is offered must not interfere with the truth-telling behavior in the good equilibrium, i.e., it must belong to the reward set corresponding to report θ_{-i}'.

Definition 8 (Refutable Deception)
A deception β is refutable if there exist i, θ_i, $\theta_i' \in \beta_i\left(\theta_i\right)$ such that, for all $\theta_{-i}' \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$, there exists y such that:

$$\sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i\left(\theta_{-i}\right) u_i\left(y, \left(\theta_i, \theta_{-i}\right)\right) > \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i\left(\theta_{-i}\right) u_i\left(f\left(\theta_i', \theta_{-i}'\right), \left(\theta_i, \theta_{-i}\right)\right), \tag{3}$$

and for all $\theta_i'' \in \Theta_i$:

$$u_i\left(f\left(\theta_i'', \theta_{-i}'\right), \left(\theta_i'', \theta_{-i}'\right)\right) \geq u_i\left(y, \left(\theta_i'', \theta_{-i}'\right)\right). \tag{4}$$

The deception is strictly refutable if, for all θ_i'' with $f\left(\theta_i'', \theta_{-i}'\right) \neq y$, the inequality (4) is strict.

This gives the following conditions:

Definition 9 (Robust Monotonicity)
Social choice function f satisfies robust monotonicity if every unacceptable deception is refutable. Social choice function f satisfies strict *robust monotonicity if every unacceptable deception is strictly refutable.*

It turns out that robust monotonicity implies the incentive compatibility conditions described above, by considering simple deceptions where a single type of one player has a single possible misreport.[9]

Lemma 1 *If f satisfies robust monotonicity, then f satisfies semi-strict EPIC.*

Proof. Suppose that $f\left(\theta_i, \theta'_{-i}\right) \neq f\left(\theta'_i, \theta'_{-i}\right)$ for some θ_i, θ'_i and θ'_{-i}. Then the deception β with

$$\beta_j\left(\widetilde{\theta}_j\right) = \begin{cases} \left\{\theta_i, \theta'_i\right\}, & \text{if} \quad \left(j, \widetilde{\theta}_j\right) = (i, \theta_i), \\ \left\{\widetilde{\theta}_j\right\} & \text{if} \qquad \text{otherwise;} \end{cases}$$

is unacceptable. Thus, by robust monotonicity, it is refutable. So for any θ_{-i}, there exists y such that

$$u_i\left(y, (\theta_i, \theta_{-i})\right) > u_i\left(f\left(\theta'_i, \theta_{-i}\right), (\theta_i, \theta_{-i})\right),$$

and for all $\theta''_i \in \Theta_i$:

$$u_i\left(f\left(\theta''_i, \theta_{-i}\right), \left(\theta''_i, \theta_{-i}\right)\right) \geq u_i\left(y, \left(\theta''_i, \theta_{-i}\right)\right).$$

Thus

$$u_i\left(f\left(\theta_i, \theta_{-i}\right), (\theta_i, \theta_{-i})\right) \geq u_i\left(y, (\theta_i, \theta_{-i})\right) > u_i\left(f\left(\theta'_i, \theta_{-i}\right), (\theta_i, \theta_{-i})\right),$$

which establishes semi-strict EPIC. ■

When we prove in Theorem 2 that robust monotonicity (together with an additional condition) is sufficient for rationalizable implementation, we will use the fact that every unacceptable deception is refutable. However, in proving the necessity of strict robust monotonicity in our next result, we will use the contrapositive statement that is closer to Maskin's original formulation of monotonicity: a social choice function satisfies strict robust monotonicity if every deception that is not strictly refutable is acceptable.[10] For this argument, it is useful to observe that a deception is not strictly refutable if, for each i, θ_i, $\theta'_i \in \beta_i(\theta_i)$, there exist $\theta'_{-i} \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta'_{-i}\right)\right)$ such that

$$u_i\left(f\left(\theta''_i, \theta'_{-i}\right), \left(\theta''_i, \theta'_{-i}\right)\right) > u_i\left(y, \left(\theta''_i, \theta'_{-i}\right)\right), \text{ for all } \theta''_i \text{ such that } f\left(\theta''_i, \theta'_{-i}\right) \neq y, \tag{5}$$

[9] This was pointed out to us by Olivier Tercieux.

[10] Discussions with Olivier Tercieux suggested this proof.

implies

$$\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)u_i\left(f\left(\theta'\right),\theta\right)\geq\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)u_i\left(y,\theta\right). \tag{6}$$

We now establish a necessary condition for rationalizable implementation.

Theorem 1 (Necessary Conditions)
If f is implementable in rationalizable strategies, then f satisfies strict robust monotonicity (and thus semi-strict EPIC).

By Lemma 1, semi-strict EPIC is also necessary. In the following proof of the Theorem, note that some extra steps are required to deal with the fact that best responses need not exist to all conjectures agents might have about others' types and messages.

Proof. Let $\mathcal{M}$ be a mechanism that implements f in rationalizable strategies. Suppose that β is not strictly refutable. Define S^β by

$$S_i^\beta\left(\theta_i\right)\triangleq\bigcup_{\theta'_i\in\beta_i(\theta_i)}S_i^{\mathcal{M}}\left(\theta'_i\right).$$

We claim that $S^\beta\leq b(S^\beta)$. To see why, fix any i, θ_i, $\theta'_i\in\beta_i\left(\theta_i\right)$. Because β is not strictly refutable (see (5) and (6)), there exist $\theta'_{-i}\in\Theta_{-i}$ and $\psi_i\in\Delta\left(\beta_{-i}^{-1}\left(\theta'_{-i}\right)\right)$ such that

$$u_i\left(f\left(\theta''_i,\theta'_{-i}\right),\left(\theta''_i,\theta'_{-i}\right)\right)>u_i\left(y,\left(\theta''_i,\theta'_{-i}\right)\right),\text{ for all }\theta''_i\text{ such that }f\left(\theta''_i,\theta'_{-i}\right)\neq y, \tag{7}$$

implies

$$\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)u_i\left(f\left(\theta'\right),\theta\right)\geq\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)u_i\left(y,\theta\right).$$

By the definition of rationalizable implementation, see Definition 4.2, there exists a belief over the messages m_{-i}, given by $\nu_i^{\theta'_{-i}}\in\Delta\left(S_{-i}^{\mathcal{M}}\left(\theta'_{-i}\right)\right)$, such that

$$\arg\max_{m_i}\sum_{m_{-i}}\nu_i^{\theta'_{-i}}\left(m_{-i}\right)u_i\left(g\left(m_i,m_{-i}\right),\left(\theta''_i,\theta'_{-i}\right)\right)\neq\varnothing, \tag{8}$$

for all $\theta''_i\in\Theta_i$. Now, for every $\widetilde{m}_i\in M_i$, define

$$y\left(\widetilde{m}_i\right)\triangleq\sum_{m_{-i}}\nu_i^{\theta'_{-i}}\left(m_{-i}\right)g\left(\widetilde{m}_i,m_{-i}\right).$$

Now if $y\left(\widetilde{m}_i\right)\neq f\left(\theta''_i,\theta'_{-i}\right)$, then we must have

$$u_i\left(f\left(\theta''_i,\theta'_{-i}\right),\left(\theta''_i,\theta'_{-i}\right)\right)>u_i\left(y\left(\widetilde{m}_i\right),\left(\theta''_i,\theta'_{-i}\right)\right),$$

for if not, then $\widetilde{m}_i$ would be rationalizable for type θ_i'', contradicting rationalizable implementation. But now, by (7),

$$\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)u_i\left(f\left(\theta'_i,\theta'_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right)\geq\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)u_i\left(y\left(\widetilde{m}_i\right),\left(\theta_i,\theta_{-i}\right)\right)$$

for all $\widetilde{m}_i$, and thus, for any $m'_i\in S_i^{\mathcal{M}}\left(\theta'_i\right)$ we have:

$$\begin{aligned}
&\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right),m_{-i}}\psi_i\left(\theta_{-i}\right)\nu_i^{\theta'_{-i}}\left(m_{-i}\right)u_i\left(g\left(m'_i,m_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right)\\
=&\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)u_i\left(f\left(\theta'_i,\theta'_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right)\\
\geq&\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)u_i\left(y\left(\widetilde{m}_i\right),\left(\theta_i,\theta_{-i}\right)\right)\\
=&\sum_{\theta_{-i}\in\beta_{-i}^{-1}\left(\theta'_{-i}\right)}\psi_i\left(\theta_{-i}\right)\nu_i^{\theta'_{-i}}\left(m_{-i}\right)u_i\left(g\left(\widetilde{m}_i,m_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right).
\end{aligned}$$

But now $S^\beta\leq b\left(S^\beta\right)\Rightarrow S^\beta\leq S^{\mathcal{M}}\Rightarrow\beta$ is acceptable. ■

We observe that the proof of strict robust monotonicity used the full strength of rationalizable implementation given in Definition 4.2. In particular, it was required that the same distribution over rationalizable messages of other agents, $\nu_i^{\theta'_{-i}}\in\Delta\left(S_{-i}^{\mathcal{M}}\left(\theta'_{-i}\right)\right)$, guaranteed a best response for every payoff type θ_i'' of agent i.

3.2 Sufficiency

We will need a very weak economic condition to ensure that it is always possible to reward and punish each agent independently of the other agents.

Definition 10 (Conditional No Total Indifference)
The conditional no total indifference (NTI) property is satisfied if, for all i, θ_i, θ'_{-i} and $\psi_i\in\Delta\left(\Theta_{-i}\right)$, there exists $y,y'\in Y_i\left(\theta'_{-i}\right)$ such that

$$\sum_{\theta_{-i}\in\Theta_{-i}}\psi_i\left(\theta_{-i}\right)u_i\left(y,\left(\theta_i,\theta_{-i}\right)\right)>\sum_{\theta_{-i}\in\Theta_{-i}}\psi_i\left(\theta_{-i}\right)u_i\left(y',\left(\theta_i,\theta_{-i}\right)\right).$$

The *conditional no total indifference property* imposes a very weak restriction on the preferences. The conditional NTI property, together with the use of lotteries, allows us to dispense with any no veto property which typically appear in the sufficient conditions. In addition, we can omit the usual cardinality assumption of $I\geq 3$. A related no total indifference condition appears in the

context of virtual implementation in Duggan (1997), who requires it to hold at every ex post profile θ and in Serrano and Vohra (2005), who require it at the interim level for a given belief $\psi_i(\theta_{-i})$ of player i.

Theorem 2 (Sufficient Conditions)
If f satisfies robust monotonicity and the conditional NTI property, then f is rationalizably implementable.

Proof. We explicitly construct the implementing mechanism. The mechanism will use "interior" lotteries over the deterministic outcome set Z and over the reward sets $Y_i(\theta_{-i})$. Given an arbitrary labelling of the outcome set $Z = \{z_0, z_1, ..., z_k, ...\}$, we define an "interior" lottery over the set Z by

$$\overline{y} \triangleq (\overline{y}_0, \overline{y}_1, ..., \overline{y}_k, ...), \tag{9}$$

where

$$\bar{y}_k \triangleq \Pr(z = z_k) = \frac{\delta^k}{1-\delta},$$

for some $\delta \in (0,1)$. For every given profile θ_{-i}, the reward set $Y_i(\theta_{-i})$ is by construction a convex set with at most a countable number of extreme points. We denote the set of extreme points of $Y_i(\theta_{-i})$ by $Y_i^*(\theta_{-i})$ and for some labelling of the points in the set we have $Y_i^*(\theta_{-i}) = \left\{y_{0,\theta_{-i}}, y_{1,\theta_{-i}}, ..., y_{l,\theta_{-i}}, ...\right\}$. An extreme point $y_{l,\theta_{-i}}$ in $Y_i^*(\theta_{-i})$ may be a deterministic or a random outcome and assigns probability $y_{l,\theta_{-i}}(z_k)$ to the pure outcome z_k. For every reward set $Y_{-i}(\theta_{-i})$, we define an "interior" lottery:

$$\bar{y}_{\theta_{-i}} = \left(\bar{y}_{0,\theta_{-i}}, \bar{y}_{1,\theta_{-i}}, ..., \bar{y}_{k,\theta_{-i}}\right) \tag{10}$$

with

$$\bar{y}_{k,\theta_{-i}} \triangleq \frac{1}{1-\delta}\sum_{l=0}^{\infty}\delta^l y_{l,\theta_{-i}}(z_k),$$

where the lottery $\bar{y}_{\theta_{-i}}$ is a compound lottery.

Each agent i sends a message $m_i = \left(m_i^1, m_i^2, m_i^3, m_i^4\right)$, where $m_i^1 \in \Theta_i$, $m_i^2 \in \mathbb{Z}_+$, $m_i^3 : \Theta_{-i} \to Y$ with $m_i^3(\theta_{-i}) \in Y_i(\theta_{-i})$, $m_i^4 \in Y$. The outcome $g(m)$ is determined by the following rules:

Rule 1: If $m_i^2 = 1$ for all i, pick $f\left(m^1\right)$.

Rule 2: If there exists $j \in I$ such that $m_i^2 = 1$ for all $i \neq j$ and $m_j^2 > 1$, then pick $m_j^3\left(m_{-j}^1\right)$ with probability $1 - 1/\left(m_j^2 + 1\right)$ and $\bar{y}_{m_{-j}^1}$ (as defined in (10)) with probability $1/\left(m_j^2 + 1\right)$.

Rule 3: In all other cases, for each i, with probability $(1/I)\left(1 - 1/\left(m_i^2 + 1\right)\right)$ pick m_i^4, and with probability $(1/I)\cdot\left(1/\left(m_i^2 + 1\right)\right)$ pick the interior lottery $\bar{y}$ (as defined in (9)).

Claim 1: It is never a best reply for type θ_i to send a message with $m_i^2 > 1$ (i.e., $m_i \in b_i\left(\overline{S}\right) \Rightarrow m_i^2 = 1$).

Proof of Claim 1: Suppose that θ_i has conjecture $\lambda_i \in \Delta\left(M_{-i} \times \Theta_{-i}\right)$. We partition the messages of other agents as follows:

$$M_{-i}^2\left(\theta_{-i}\right) = \left\{m_{-i} : m_j^2 = 1 \text{ for all } j \neq i \text{ and } m_{-i}^1 = \theta_{-i}\right\},$$

and

$$M_{-i}^3 = \left\{m_{-i} : m_j^2 > 1 \text{ for some } j \neq i\right\}.$$

The set $M_{-i}^2\left(\theta_{-i}\right)$ is the set of messages such that either Rule 1 or Rule 2 is triggered. The set M_{-i}^3 is such that either Rule 2 or Rule 3 is triggered. The notation reminds us that if agent i chooses $m_i^2 > 1$, then $M_{-i}^2\left(\theta_{-i}\right)$ and M_{-i}^3 trigger Rule 2 and Rule 3, respectively. By the conditional NTI property, we know that there exists $m_i^4 \in Y$ such that, if

$$\sum_{m_{-i} \in M_{-i}^3, \theta_{-i} \in \Theta_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) > 0,$$

then

$$\sum_{m_{-i} \in M_{-i}^3, \theta_{-i} \in \Theta_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) u_i\left(m_i^4, \theta\right) > \sum_{m_{-i} \in M_{-i}^3, \theta_{-i} \in \Theta_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) u_i\left(\bar{y}, \theta\right).$$

And we also know from the conditional NTI property that there exists m_i^3 such that, if

$$\sum_{m_{-i} \in M_{-i}^2\left(\theta'_{-i}\right), \theta_{-i} \in \Theta_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) > 0,$$

then

$$\sum_{m_{-i} \in M_{-i}^2\left(\theta'_{-i}\right), \theta_{-i} \in \Theta_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) u_i\left(m_i^3\left(\theta'_{-i}\right), \theta\right) > \sum_{m_{-i} \in M_{-i}^2\left(\theta'_{-i}\right), \theta_{-i} \in \Theta_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) u_i\left(\bar{y}_{\theta_{-i}}, \theta\right).$$

Thus if $\left(m_i^1, m_i^2, m_i^3, m_i^4\right)$ with $m_i^2 > 1$ were a best response, then $\left(m_i^1, m_i^2 + 1, m_i^3, m_i^4\right)$ would be an even better response, a contradiction.

Claim 2: $\left(\theta'_i, m_i^2, m_i^3, m_i^4\right) \in S_i\left(\theta_i\right) \Rightarrow f\left(\theta'_i, \theta_{-i}\right) = f\left(\theta_i, \theta_{-i}\right)$ for all $\theta_{-i} \in \Theta_{-i}$.

Proof of Claim 2: Now fix any S with $m_i \in S_i\left(\theta_i\right) \Rightarrow m_i^2 = 1$. Let

$$\beta_i\left(\theta_i\right) = \left\{\theta'_i : \left(\theta'_i, 1, m_i^3, m_i^4\right) \in S_i\left(\theta_i\right) \text{ for some } \left(m_i^3, m_i^4\right)\right\}.$$

We will argue that if β is not acceptable, then $b\left(S\right) \neq S$. By robust monotonicity, we know that there exists i, θ_i, $\theta'_i \in \beta_i\left(\theta_i\right)$ such that, for all $\theta'_{-i} \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta'_{-i}\right)\right)$, there exists $y \in Y_i\left(\theta'_{-i}\right)$ such that

$$\sum_{\theta_{-i} \in \Theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(y, \left(\theta_i, \theta_{-i}\right)\right) > \sum_{\theta_{-i} \in \Theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(f\left(\theta'_i, \theta'_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right).$$

But now for any conjecture $\lambda_i \in \Delta\left(\left\{(m_{-i},\theta_{-i}) : m_j^2 = 1 \text{ for all } j \neq i\right\}\right)$, there exists m_i^3 (with $m_i^3(\theta_{-i}) \in Y_i(\theta_{-i})$) such that

$$\sum_{m_{-i},\theta_{-i}} \lambda_i(m_{-i},\theta_{-i})\, u_i\left(m_i^3\left(m_{-i}^1\right),\theta\right) > \sum_{m_{-i},\theta_{-i}} \lambda_i(m_{-i},\theta_{-i})\, u_i\left(f\left(\theta_i',m_{-i}^1\right),(\theta_i,\theta_{-i})\right).$$

Thus message $\left(\theta_i',1,m_i^3,m_i^4\right)$ is never a best response for type θ_i. We conclude that if

$$\beta_i(\theta_i) = \left\{\theta_i' : \left(\theta_i',1,m_i^3,m_i^4\right) \in S_i^{\mathcal{M}}(\theta_i) \text{ for some } \left(m_i^3,m_i^4\right)\right\},$$

then β is acceptable.

Claim 3: $\left(\theta_i,1,m_i^3,m_i^4\right) \in S_i^{\mathcal{M}}(\theta_i)$ for all i and θ_i.

Proof of Claim 3: This is an immediate consequence of EPIC, which (by Lemma 1) is implied by strict robust monotonicity.

It now follows from Claims 1-3 that f is rationalizably implemented. ■

Observe that since robust monotonicity and conditional NTI are sufficient for rationalizable implementation, and strict robust monotonicity is necessary for rationalizable implementation, it follows that robust monotonicity and conditional NTI must imply strict robust monotonicity. It is straightforward to check this directly.

We allowed for badly behaved infinite mechanisms in order to make a tight connection with the existing literature and to get tight results. Many authors have argued that "integer game" constructions, like the one used in the above theorem, should be viewed with suspicion (see, e.g., Abreu and Matsushima (1992a) and Jackson (1992)).

3.3 A Coordination Example

We conclude this section with an example that demonstrates that while rationalizable implementation is a strong requirement, it is weaker than dominant strategy implementation. In the example there are two agents, $i = 1, 2$. Each agent i has two possible types, θ_i and θ_i'. There are six possible outcomes: $Z = \{a, b, c, d, z, z'\}$. The payoffs of the agents are a function of the allocation and the true payoff type profile, given by:

a	θ_2	θ_2'
θ_1	$3,3$	$0,0$
θ_1'	$0,0$	$1,1$

b	θ_2	θ_2'
θ_1	$0,0$	$3,3$
θ_1'	$1,1$	$0,0$

c	θ_2	θ_2'
θ_1	$0,0$	$1,1$
θ_1'	$3,3$	$0,0$

d	θ_2	θ_2'
θ_1	$1,1$	$0,0$
θ_1'	$0,0$	$3,3$

(11)

and

z	θ_2	θ_2'
θ_1	$2,2$	$2,0$
θ_1'	$2,2$	$2,0$

z′	θ_2	θ_2'
θ_1	$2,0$	$2,2$
θ_1'	$2,0$	$2,2$

.

The social choice function is given by the efficient outcome at each type profile:

f	θ_2	θ_2'
θ_1	a	b
θ_1'	c	d

. (12)

Clearly, the social choice function is strictly ex post incentive compatible. But in the "direct mechanism" where each agent simply reports his type, there will always be an equilibrium where each type of each agent misreports his type, and each agent gets a payoff of 1. This is also strictly ex post incentive compatible. The social choice function f which selects among $\{a, b, c, d\}$ embeds a coordination game. We further observe that the payoff for agent 1 from allocations z and z' are equal and constant for all type profiles. On the other hand, the payoff of agent 2 from z and z' depends on his type but not on the type of the other agent.

We now consider the following augmented, but finite, mechanism which responds to the messages of the agents as follows:

g	θ_2	θ_2'
θ_1	a	b
θ_1'	c	d
ζ	z	z'

.

The augmented mechanism enriches the message space of agent 1 by a single message ζ. The corresponding incomplete information game has the following payoffs:

	type	θ_2		θ_2'	
type	action	θ_2	θ_2'	θ_2	θ_2'
θ_1	θ_1	$3,3$	$0,0$	$0,0$	$3,3$
	θ_1'	$0,0$	$1,1$	$1,1$	$0,0$
	ζ	$2,2$	$2,0$	$2,0$	$2,2$
θ_1'	θ_1	$0,0$	$1,1$	$1,1$	$0,0$
	θ_1'	$3,3$	$0,0$	$0,0$	$3,3$
	ζ	$2,2$	$2,0$	$2,0$	$2,2$

.

Suppose we iteratively remove actions for each type that could never be a best response given the type action profiles remaining. Thus in the first round, we would observe that type θ_1 would never send message θ_1' and type θ_1' would never send message θ_1. Knowing this, we could conclude that type θ_2 would never send message θ_2' and type θ_2' would never send message θ_2. This in turn implies that neither type of agent 1 will ever send message ζ. Thus the only remaining message for each type of each agent is truth-telling. But now they must behave this way in any equilibrium on any type space.

4 Robust Implementation

We now establish the equivalence between rationalizable and robust implementation. We first report a formal epistemic argument that relates the rationalizable messages to the set of messages that might be played in any equilibrium on any type space. We use the notation of type spaces that we introduced in Section 2.2.

Proposition 1 (Rationalizable Actions)
$m_i \in S_i^{\mathcal{M}}(\theta_i)$ if and only if there exists a type space $\mathcal{T}$, an interim equilibrium σ of the game $(\mathcal{T}, \mathcal{M})$ and a type $t_i \in T_i$ such that (i) $\sigma_i(m_i|t_i) > 0$ and (ii) $\widehat{\theta}_i(t_i) = \theta_i$.

Proof. ($\Rightarrow$) Suppose $m_i^* \in S^{\mathcal{M}}(\theta_i^*)$. Now consider the following type space $\mathcal{T}$ defined through $T_i \triangleq \left\{(m_i, \theta_i) \left| m_i \in S_i^{\mathcal{M}}(\theta_i) \right.\right\}$ and let $\widehat{\theta}_i(m_i, \theta_i) \triangleq \theta_i$. By the definition of rationalizability, we know that for each $m_i \in S_i^{\mathcal{M}}(\theta_i)$, there exists $\lambda_i^{m_i,\theta_i} \in \Delta(M_{-i} \times \Theta_{-i})$ such that $\lambda_i^{m_i,\theta_i}(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j^{\mathcal{M}}(\theta_j)$ for each $j \neq i$; and

$$\sum_{m_{-i},\theta_{-i}} \lambda_i^{m_i,\theta_i}(m_{-i}, \theta_{-i}) \left[u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i})) - u_i\left(g\left(m_i', m_{-i}\right), (\theta_i, \theta_{-i})\right)\right] \geq 0, \ \forall m_i' \in M_i.$$

Let $\widehat{\pi}_i(m_{-i}, \theta_{-i})[m_i, \theta_i] \triangleq \lambda_i^{m_i,\theta_i}(m_{-i}, \theta_{-i})$. Now by construction, there is a pure strategy equilibrium s with $s_i(m_i, \theta_i) = m_i$. But now $s_i(m_i^*, \theta_i^*) = m_i^*$ and $\widehat{\theta}(m_i^*, \theta_i^*) = \theta_i^*$.

($\Leftarrow$) Suppose there exists a type space $\mathcal{T}$, an equilibrium σ of $(\mathcal{T}, \mathcal{M})$, and $m_i^* \in M_i$ and $t_i^* \in T_i$ such that $\sigma_i(m_i^*|t_i^*) > 0$ and $\widehat{\theta}_i(t_i^*) = \theta_i^*$. Let

$$S_i(\theta_i) = \left\{m_i : \sigma_i(m_i|t_i) > 0 \text{ and } \widehat{\theta}_i(t_i) = \theta_i \text{ for some } t_i \in T_i\right\}.$$

Now interim equilibrium conditions ensure that $b(S) \geq S$. Thus $S \leq S^{\mathcal{M}}$. Thus $m_i^* \in S_i^{\mathcal{M}}\left(\widehat{\theta}_i(t_i^*)\right)$, which concludes the proof. ■

We emphasize that the above equivalence result does not guarantee that there exist an equilibrium on every type space. Since there may be an infinite number of messages, the existence of an equilibrium is in fact not guaranteed. We will use the following condition in establishing the existence of interim equilibria on all type spaces. The condition uses the notion of message correspondence S defined in Section 2.4.

Definition 11 (Ex Post Best Response)
Message correspondence S satisfies the ex post best response property if, for all i and $\theta_i \in \Theta_i$, there exists $m_i^* \in S_i(\theta_i)$ such that:

$$m_i^* \in \arg\max_{m_i \in M_i} u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i})),$$

for all θ_{-i} and $m_{-i} \in S_{-i}(\theta_{-i})$.

This condition requires that for each i and θ_i, there is a single message which is rationalizable whatever agent i's beliefs about others' payoff types.

Theorem 3 (Almost Equivalence)

1. *If f is rationalizably implementable by mechanism $\mathcal{M}$, and $S^{\mathcal{M}}$ satisfies the ex post best response property, then f is robustly implementable by mechanism $\mathcal{M}$.*

2. *If f is robustly implementable by mechanism $\mathcal{M}$, then f is rationalizably implementable by mechanism $\mathcal{M}$.*

Proof. (1.) By the ex post best response property, there exists, for each i, $s_i^* : \Theta_i \to M_i$ such that

$$s_i^*(\theta_i) \in \arg\max_{m_i \in M_i} u_i\left(g\left(m_i, s_{-i}^*(\theta_{-i})\right), (\theta_i, \theta_{-i})\right)$$

for all θ_{-i}. Now fix any type space. The strategy profile s with $s_i(t_i) = s_i^*(\widehat{\theta}_i(t_i))$ is an equilibrium of the game $(\mathcal{T}, \mathcal{M})$. Thus *an* interim equilibrium $s(t)$ with the property that $g(s(t)) = f\left(\widehat{\theta}(t)\right)$ exist for every type space $\mathcal{T}$. Now by Proposition 1, every equilibrium action m_i which is chosen with positive probability, $\sigma_i(m_i|t_i) > 0$, is also a rationalizable action.

(2). Suppose $(\mathcal{T}, \mathcal{M})$ has an equilibrium for each $\mathcal{T}$. Fix any i and $\psi_i \in \Delta(\Theta_{-i})$. Fix any type space $\mathcal{T}$ with, for each $\theta_i \in \Theta_i$, a type $t_i^*(\theta_i)$ such that (a) $\widehat{\theta}_i(t_i^*(\theta_i)) = \theta_i$ for each θ_i, (b) there exists $\pi_i \in \Delta(T_{-i})$ such that $\widehat{\pi}_i(t_i^*(\theta_i)) = \pi_i$ for all θ_i and (c)

$$\sum_{\left\{t_{-i}:\widehat{\theta}_{-i}(t_{-i})=\theta_{-i}\right\}} \pi_i(t_{-i}) = \psi_i(\theta_{-i}) \tag{13}$$

for all θ_i and θ_{-i}. The game has an equilibrium σ. Let m_i be any message with $\sigma_i(m_i|t_i^*(\theta_i)) > 0$. Let

$$\lambda_i(m_{-i}, \theta_{-i}) = \sum_{\left\{t_{-i} \in T_{-i}:\widehat{\theta}_{-i}(t_{-i})=\theta_{-i}\right\}} \sigma_{-i}(m_{-i}|t_{-i})\,\pi_i(t_{-i}).$$

Now $\sigma_i(m_i|t_i^*(\theta_i)) > 0$ implies

$$m_i(\theta_i) \in \arg\max_{m_i \in M_i} \sum_{m_{-i}, \theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i})).$$

Proposition 1 implies that every message profile m_j which is played in equilibrium by type θ_j is part of the set $S^{\mathcal{M}}$, or that:

$$\lambda_i(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j^{\mathcal{M}}(\theta_j) \text{ for each } j \neq i.$$

By construction of the type space $\mathcal{T}$, in particular property (c) as expressed by (13), this implies that

$$\sum_{m_{-i} \in M_{-i}} \lambda_i \left(m_{-i}, \theta_{-i} \right) = \psi_i \left(\theta_{-i} \right) \text{ for all } \theta_{-i} \in \Theta_{-i}.$$

Since these properties hold for arbitrary i and $\psi_i \in \Delta \left(\Theta_{-i} \right)$, part (2) of the definition of rationalizable implementation is satisfied. ■

It is unfortunate that there is a gap in the above proposition. However, an example in the appendix shows that it is possible to construct (admittedly silly) mechanisms where $(\mathcal{T}, \mathcal{M})$ has an equilibrium for each $\mathcal{T}$, but $S^{\mathcal{M}}$ fails the ex post best response property.

Corollary 1

1. *If f is robustly implementable, then f satisfies EPIC and strict robust monotonicity;*
2. *If f satisfies EPIC, robust monotonicity and the conditional NTI property, then f is robustly implementable.*

Proof. (1.) It follows immediately from Theorem 3.2 that if f robustly implementable then it is also rationalizably implementable, and if follows from Theorem 1 that if f is rationalizably implementable, then it satisfies EPIC and strict robust monotonicity.

(2.) By Theorem 2, if f satisfies EPIC, robust monotonicity and the conditional NTI property, then f is rationalizably implementable. Now since f satisfies EPIC, we observe that $S^{\mathcal{M}}$ must satisfy the ex post best response property, with type θ_i sending a message of the form $\left(\theta_i, 1, m_i^3, m_i^4 \right)$, and so robust implementation is possible by Theorem 3.1. ■

5 Finite Mechanisms

In this section, we restrict attention to finite mechanisms, i.e. where each M_i is finite. All the results in this section extend to compact or, more generally, "regular" mechanisms (e.g., mechanisms where best responses always exist as in Abreu and Matsushima (1992b)). We present a different necessary condition - robust measurability - that arises when attention is restricted to finite mechanisms.[11] We can also use the case of finite mechanisms to discuss the relationship with some of our earlier work providing sufficient conditions for implementation using well behaved mechanisms by (i) restricting attention to more restrictive environments (Bergemann and Morris (2009a)) or (ii) allowing virtual implementation (Bergemann and Morris (2009b)).

[11] While the stronger necessary conditions apply even if we allow infinite payoff type spaces, clearly only very restrictive social choice functions can be implemented by finite mechanisms in this case.

5.1 Necessary Condition

We report an additional necessary condition (from Bergemann and Morris (2009b)) for robust implementation in finite mechanisms. We are interested in the set of preferences that an agent might have if his payoff type is θ_i and he knows that the type θ_j of each opponent j belongs to some subset Ψ_j of his payoff types Θ_j. Write $\mathcal{R}$ for the set of expected utility preference relations on lotteries Y. We will write $R_{\theta_i,\psi_i} \in \mathcal{R}$ for the preference relation of agent i if his payoff type is θ_i and he has belief $\psi_i \in \Delta(\Theta_{-i})$ about the types of others:

$$\forall y, y' \in Y: \qquad yR_{\theta_i,\psi_i}y' \quad \Leftrightarrow \quad \sum_{\theta_{-i}\in\Theta_{-i}} \psi_i(\theta_{-i})\, u_i(y,(\theta_i,\theta_{-i})) \geq \sum_{\theta_{-i}\in\Theta_{-i}} \psi_i(\theta_{-i})\, u_i\left(y',(\theta_i,\theta_{-i})\right).$$

We write $\mathcal{R}_i(\theta_i, \Psi_{-i})$ for the set of preferences agent i might have if his payoff type is θ_i and he might have any beliefs over others' payoff types:

$$\mathcal{R}_i(\theta_i,\Psi_{-i}) \triangleq \left\{R \in \mathcal{R} \,\middle|\, R = R_{\theta_i,\psi_i} \text{ for some } \psi_i \in \Delta(\Psi_{-i})\right\}.$$

Say that type set profile Ψ_{-i} *separates* Ψ_i if

$$\bigcap_{\theta_i\in\Psi_i} \mathcal{R}_i(\theta_i,\Psi_{-i}) = \varnothing.$$

Let $\Xi = (\Xi_i)_{i=1}^I \subseteq \times_{i=1}^I 2^{\Theta_i}$ be a profile of type sets for each agent. Say that Ξ is *mutually inseparable* if, for each i and $\Psi_i \in \Xi_i$, there exists $\Psi_{-i} \in \Xi_{-i}$ such that Ψ_{-i} does not separate Ψ_i.

Definition 12 (Robust Measurability)
Social choice function f satisfies robust measurability if Ξ mutually inseparable, $\Psi_i \in \Xi_i$ and $\left\{\theta_i', \theta_i''\right\} \subseteq \Psi_i \Rightarrow f\left(\theta_i', \theta_{-i}\right) = f\left(\theta_i'', \theta_{-i}\right)$ for all θ_{-i}.

If payoff types are finite, one can give an alternative iterative definition of robust measurability: let $\Xi_i^0 = 2^{\Theta_i}$, and inductively define:

$$\Xi_i^{k+1} = \left\{\Psi_i \in \Xi_i^k \mid \Psi_{-i} \text{ does not separate } \Psi_i, \text{ for some } \Psi_{-i} \in \Xi_{-i}^k\right\},$$

and

$$\Xi_i^* = \bigcap_{k\geq 0} \Xi_i^k.$$

Now we say that a social choice function f satisfies robust measurability if

$$\left\{\theta_i', \theta_i''\right\} \in \Xi_i^* \quad \Rightarrow \quad f\left(\theta_i', \theta_{-i}\right) = f\left(\theta_i'', \theta_{-i}\right) \quad \text{for all } \theta_{-i}.$$

Bergemann and Morris (2009b) showed that robust measurability was necessary for robust *virtual* implementation; thus it must also be necessary of robust exact implementation. For completeness, we report a direct argument here.

Proposition 2 (Necessity of Robust Measurability)
If social choice function f is robustly implementable by a finite mechanism, then f satisfies robust measurability.

Proof. Since f is robustly implementable, there exists a mechanism $\mathcal{M}$ such that

$$m \in S^{\mathcal{M}}(\theta) \Rightarrow g(m) = f(\theta).$$

Now suppose Ξ is mutually inseparable. We argue by induction that, for all i, $\Psi_i \in \Xi_i$ and k there exists a set of messages $\varnothing \neq M_i^k(\Psi_i) \subseteq S_i^{\mathcal{M},k}(\theta_i)$ for all $\theta_i \in \Psi_i$. This is true by definition for $k = 0$. Suppose that it is true for k. Now Ξ being mutually inseparable implies that for any $\Psi_i \in \Xi_i$, there exists $\Psi_{-i} \in \Xi_{-i}$, R and, for each $\theta_i \in \Psi_i$, $\lambda_i^{\theta_i} \in \Delta(\Psi_{-i})$ such that $R_{\theta_i, \lambda_i^{\theta_i}} = R$. Now let $M_i^{k+1}(\Psi_i)$ be the optimal messages of agent i when he believes that his opponents will sent some message profile in $M_{-i}^k(\Psi_{-i})$ with probability 1 and has beliefs $\lambda_i^{\theta_i}$ about the type profile of his opponents, i.e.,

$$M_i^{k+1}(\Psi_i) = \bigcup_{m_{-i} \in M_{-i}^k(\Psi_{-i})} \arg\max_{m_i'} \sum_{\theta_{-i}} \lambda_i^{\theta_i}(\theta_{-i}) u_i\left(g\left(m_i', m_{-i}\right), (\theta_i, \theta_{-i})\right).$$

By construction, $\varnothing \neq M_i^{k+1}(\Psi_i) \subseteq S_i^{\mathcal{M},k+1}(\theta_i)$ for all $\theta_i \in \Psi_i$. Now for each $\Psi_i \in \Xi_i$, $M_i^k(\Psi_i)$ is a decreasing sequence under set inclusion. Since M_i is finite, there exists

$$M_i^*(\Psi_i) = \bigcap_{k \geq 0} M_i^k(\Psi_i) \neq \varnothing.$$

Thus $M_i^*(\Psi_i) \subseteq S_i^{\mathcal{M}}(\theta_i)$ for all $\theta_i \in \Psi_i$. Now if $\left\{\theta_i', \theta_i''\right\} \subseteq \Psi_i$, there exists $m_i \in M_i^*(\Psi_i) \subseteq S_i^{\mathcal{M}}\left(\theta_i'\right)$ and $m_i \in M_i^*(\Psi_i) \subseteq S_i^{\mathcal{M}}\left(\theta_i''\right)$. Now fix any $m_{-i} \in S_{-i}^{\mathcal{M}}(\theta_{-i})$, and we have $(m_i, m_{-i}) \in S^{\mathcal{M}}\left(\theta_i', \theta_{-i}\right) \Rightarrow g(m_i, m_{-i}) = f\left(\theta_i', \theta_{-i}\right)$ and $(m_i, m_{-i}) \in S^{\mathcal{M}}\left(\theta_i'', \theta_{-i}\right) \Rightarrow g(m_i, m_{-i}) = f\left(\theta_i'', \theta_{-i}\right)$. Thus $f\left(\theta_i', \theta_{-i}\right) = f\left(\theta_i'', \theta_{-i}\right)$. ■

In the appendix of the working paper version, Bergemann and Morris (2008b), we show by means of two examples that robust monotonicity does not imply nor is it implied by robust measurability.

5.2 Sufficient Conditions

We have pursued two ways of deriving sufficient conditions in prior work. First, we showed that if we weaken the implementation requirement to virtual implementation (Bergemann and Morris (2009b)), then the above robust measurability condition is sufficient (under weak conditions ruling out indifference). Second, we identified natural restrictions on the environment that make the necessary conditions sufficient (Bergemann and Morris (2009a)); we briefly review this result below.

If we neither put restrictions on the environment nor allow virtual implementation, then we do not know how to derive tight sufficient conditions for finite, or other well-behaved, mechanisms. However, as in the existing literature on complete and incomplete information implementation, it was possible to obtain tight conditions only if we allow for badly behaved mechanisms.

5.3 Single Crossing Aggregator Environments

In Bergemann and Morris (2009a), we consider payoff environments in which each payoff type space Θ_i is completely ordered and where there exist for each i, an aggregator function $h_i : \Theta \to \mathbb{R}$ and a valuation function $v_i : Y \times \mathbb{R} \to \mathbb{R}$ such that

$$v_i\left(y, h_i\left(\theta\right)\right) \triangleq u_i\left(y, \theta\right), \tag{14}$$

where h_i is continuous and strictly increasing in θ_i and $v_i : Y \times \mathbb{R} \to \mathbb{R}$ is continuous and satisfies the following strict single crossing property: for all $\phi < \phi' < \phi''$,

$$v_i\left(y, \phi\right) > v_i\left(y', \phi\right) \text{ and } v_i\left(y, \phi'\right) = v_i\left(y', \phi'\right) \quad \Rightarrow \quad v_i\left(y, \phi''\right) < v_i\left(y', \phi''\right). \tag{15}$$

The aggregator functions $h = \left(h_i\right)_{i=1}^I$ are said to satisfy the *contraction property* if, for all deceptions $\beta \neq \beta^*$, there exists i, θ_i and $\theta_i' \in \beta_i\left(\theta_i\right)$ with $\theta_i' \neq \theta_i$, such that

$$\text{sign}\left(\theta_i - \theta_i'\right) = \text{sign}\left(h_i\left(\theta_i, \theta_{-i}\right) - h_i\left(\theta_i', \theta_{-i}'\right)\right)$$

for all θ_{-i} and $\theta_{-i}' \in \beta_{-i}\left(\theta_{-i}\right)$. In single crossing aggregator environments as described by (14) and (15), the contraction property is equivalent to both strict robust monotonicity and robust measurability.

We say that a social choice function f is *responsive* if for all $\theta_i \neq \theta_i'$, there exists θ_{-i} such that $f\left(\theta_i, \theta_{-i}\right) \neq f\left(\theta_i', \theta_{-i}\right)$.

Proposition 3 (Contraction Property)
In a single crossing aggregator environment, a responsive social choice function f is robustly implementable if and only if it satisfies strict ex post incentive compatibility and the contraction property.

This result is reported in Theorem 1 and 2 of Bergemann and Morris (2009a). It follows that the necessary conditions of Theorem 1 are also sufficient in these environments. Note that in the discrete type setting of this paper, the continuity properties are automatically satisfied if the payoff type spaces are finite. Bergemann and Morris (2009a) allowed for compact payoff type spaces and pure outcome spaces; they also showed that when robust implementation is possible, it is possible in a "direct" mechanism where agents report just their payoff types.

6 Extensions, Variations and Discussion

6.1 Responsive Social Choice Functions

In Bergemann and Morris (2009a), we considered a less general environment and also that the social choice function was *responsive* in the following sense:

Definition 13 (Responsive Social Choice Function)
Social choice function f is responsive if, for all i and $\theta_i' \neq \theta_i$, there exists θ_{-i} such that $f(\theta_i, \theta_{-i}) \neq f(\theta_i', \theta_{-i})$.

Thus a social choice function is said to be responsive if distinct payoff types lead to distinct outcomes under the social choice function. Because the objective of this paper was to investigate general environments and social choice functions, we did not assume responsiveness here. However, we briefly note in this Subsection how some of our results - in particular, the necessity arguments - can be strengthened with responsiveness but in the general payoff environment of this paper. First, semi-strict EPIC immediately implies strict EPIC:

Definition 14 (Strict EPIC)
Social choice function f satisfies strict ex post incentive compatibility (strict EPIC) if, for each i, $\theta_i \neq \theta_i'$:

$$u_i(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})) > u_i\left(f\left(\theta_i', \theta_{-i}\right), (\theta_i, \theta_{-i})\right), \quad \forall \theta_{-i}.$$

Second, under strict EPIC, strict robust monotonicity is implied by the following weakening:

Definition 15 (Strict Pairwise Robust Monotonicity)
A social choice function f satisfies strict pairwise robust monotonicity if, for every unacceptable deception β, there exist i, θ_i, $\theta_i' \in \beta_i(\theta_i)$ such that, for all $\theta_{-i}' \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$, there exists y such that:

$$\sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i(\theta_{-i}) u_i(y, (\theta_i, \theta_{-i})) > \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i(\theta_{-i}) u_i\left(f\left(\theta_i', \theta_{-i}'\right), (\theta_i, \theta_{-i})\right),$$

and:

$$u_i\left(f\left(\theta_i', \theta_{-i}'\right), \left(\theta_i', \theta_{-i}'\right)\right) > u_i\left(y, \left(\theta_i', \theta_{-i}'\right)\right). \tag{16}$$

If we required condition (16) is to hold not only for θ_i', but also for any θ_i'', then this condition becomes strict robust monotonicity. We label it "pairwise" monotonicity because it involves only pairwise comparisons of types.

Finally, for responsive social choice functions, we can also establish necessary conditions under a less stringent definition of rationalizable implementation. In particular, we can weaken the requirement (c) from Definition 4.2 of rationalizable implementation by dropping the uniformity condition: "for all $\theta_i \in \Theta_i$" and instead can allow the belief $\lambda_i \in \Delta(M_{-i} \times \Theta_{-i})$ to depend on the type θ_i. We can also establish the necessity of semi-strict EPIC and strict pairwise robust monotonicity under this weakened notion, and, as noted above, with responsive social choice functions, the former implies strict EPIC and strict EPIC and the latter imply strict robust monotonicity. The proofs of the claims in this subsection are in the appendix. These result are analogues of complete information results developed in Bergemann, Morris, and Tercieux (2010) and we are grateful to Olivier Tercieux for suggesting that we pursue the implications of responsive social choice functions in this setting.

6.2 Finite Outcomes and States and Duality

We say that an environment is finite if pure outcome space Z and payoff type spaces Θ_i are all finite. We can give a simpler characterization of when a deception is refutable and thus when a social choice function satisfies robust monotonicity in finite environments:

Lemma 2 (Refutable Deception)
In finite environments, a deception β is refutable if there exist i, θ_i, $\theta_i' \in \beta_i(\theta_i)$ such that, for all $\theta_{-i}' \in \Theta_{-i}$, there exists $y \in Y$ such that for all $\theta_{-i} \in \beta_{-i}^{-1}(\theta_{-i}')$:

$$u_i(y, (\theta_i, \theta_{-i})) > u_i\left(f\left(\theta_i', \theta_{-i}'\right), (\theta_i, \theta_{-i})\right), \tag{17}$$

and for all $\theta_i'' \in \Theta_i$:

$$u_i\left(f\left(\theta_i'', \theta_{-i}'\right), \left(\theta_i'', \theta_{-i}'\right)\right) \geq u_i\left(y, \left(\theta_i'', \theta_{-i}'\right)\right). \tag{18}$$

The deception is strictly refutable if, for all θ_i'' with $f\left(\theta_i'', \theta_{-i}'\right) \neq y$, the inequality (18) is strict.

We can also give a analogous simpler characterization of conditional NTI:

Lemma 3 (Conditional No Total Indifference)
In finite environments, the conditional no total indifference (NTI) property is satisfied if, for all i, θ_i, θ_{-i}', there exists $y, y' \in Y_i\left(\theta_{-i}'\right)$ such that $u_i(y, (\theta_i, \theta_{-i})) > u_i(y', (\theta_i, \theta_{-i}))$ for all $\theta_{-i} \in \Theta_{-i}$.

These lemmas are proved the appendix. By using a duality argument, the finite versions of refutable deception and conditional no total indifference can be stated pointwise for every profile (θ_i, θ_{-i}) rather than in expectations using distributions $\psi_i(\theta_{-i}) \in \Delta(\Theta_{-i})$.

6.3 Lotteries, Pure Strategies and Bayesian Implementation

Here we discuss how Theorem 1 is related to the classic literature on Bayesian implementation developed by Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989) and Jackson (1991). These authors asked whether it was possible to implement a social choice function in equilibrium on a fixed type space $\mathcal{T}$.[12] These authors analyzed the classic problem where attention was restricted to pure strategy equilibria and deterministic mechanisms. The assumption entails that the social choice function is a mapping $f : \Theta \to Z$ and the mechanism $g : M \to Z$. Note that in this classical approach it was not necessary to even define agent's preferences over lotteries.

Having fixed a type space, the natural notion of a pure strategy deception on the fixed type space is a collection $\alpha = (\alpha_1, ..., \alpha_I)$, with each $\alpha_i : T_i \to T_i$. Thus $\alpha : T \to T$ is defined by $\alpha(t) = (\alpha_i(t_i))_{i=1}^I$. In this section we maintain that the payoff type space Θ_i and the pure outcome space Z are finite. The key monotonicity notion, translated into our language, is then the following:

Definition 16 (Bayesian Monotonicity)
Social choice function f satisfies Bayesian monotonicity on type space $\mathcal{T}$ if, for every deception α with $f\left(\widehat{\theta}(t)\right) \neq f\left(\widehat{\theta}(\alpha(t))\right)$ for some t, there exists i, t_i and $k : T \to Z$ such that

$$\sum_{t_{-i}} u_i\left(k(\alpha(t)), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_{-i})[t_i] > \sum_{t_{-i}} u_i\left(f\left(\widehat{\theta}(\alpha(t))\right), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_{-i})[t_i],$$

and

$$\sum_{t_{-i}} u_i\left(f\left(\widehat{\theta}(t_i', t_{-i})\right), \widehat{\theta}(t_i', t_{-i})\right) \widehat{\pi}_i(t_{-i})[t_i'] \geq \sum_{t_{-i}} u_i\left(k(\alpha_i(t_i), t_{-i}), \widehat{\theta}(t_i', t_{-i})\right) \widehat{\pi}_i(t_{-i})[t_i'], \ \forall t_i'.$$

Jackson (1991) shows that this condition is necessary for Bayesian implementation, and that a slight strengthening, Bayesian monotonicity no veto, is sufficient. We can also show that our robust monotonicity condition is equivalent to the requirement that Bayesian monotonicity is satisfied on all type spaces.

Theorem 4 (Equivalence)
Social choice function f satisfies Bayesian monotonicity on every type space if and only if it satisfies robust monotonicity.

The equivalence is established by a constructive proof via a specific type space. The constructive element is the identification of a type space on which Bayesian monotonicity is guaranteed to fail if robust monotonicity fails. The specific type space is much smaller than the universal type space.

[12] They allowed for more general social choice sets, but we restrict attention to functions for our comparison.

The proof of this result is in the appendix of the working paper version, Bergemann and Morris (2008b).

The notion of robustness is more subtle in the context of full rather than partial implementation. With partial implementation, i.e. truthtelling in the direct mechanism, the universal type space is by definition the most difficult type space to obtain truthtelling. In the universal type space, every agent has the maximal number of possible misreports and hence the designer faces the maximal number of incentive constraints. In the context of full implementation, the trade-off is ambiguous. As a larger type space contains by definition more types, it offers every agent more possibilities to misreport. But then, just as a larger type space made truthtelling more difficult to obtain, the other equilibria might also cease to exist after the introduction of additional types. This second part offers the possibility that larger type spaces facilitate rather than complicate the full implementation problem.

But note that this line of argument would establish the necessity of robust implementation if the planner is restricted to deterministic mechanisms (a disadvantage) but he can assume that agents follow pure strategies (an advantage). How do these assumptions matter?

First, observe that the advantage of restricting attention to pure strategies goes away completely when we require implementation on all type spaces: if there is a mixed strategy equilibrium that results in a socially sub-optimal outcome on some type space, we can immediately construct a larger type space (purifying the original equilibrium) where the socially sub-optimal outcome is played in a pure strategy equilibrium. Thus our robust analysis conveniently removes that unfortunate gap between pure and mixed strategy implementation that has plagued the implementation literature.

We use the extension to stochastic mechanisms in just two places. Ex post incentive compatibility and robust monotonicity would remain necessary conditions even if we restricted attention to deterministic mechanisms (the arguments would be unchanged). If lotteries were not allowed, our sufficiency argument would require a strengthened version of the robust monotonicity condition, with the lottery y replaced by a deterministic outcome. Our sufficiency argument also uses lotteries under Rules 1 and 2. As in recent papers by Benoit and Ok (2008) and Bochet (2007) on complete information implementation, we use lotteries to significantly weaken the sufficient conditions, so that we require only the conditional NTI property in addition to EPIC and robust monotonicity. If we did not allow lotteries in this part of the argument, we would require a much stronger economic condition in the spirit of Jackson's "Bayesian monotonicity no veto" condition. We have developed combined robust monotonicity and economic conditions (not reported here) sufficient for interim implementation on all full support types spaces. However, an additional complication is that, without lotteries in the implementing mechanism, we cannot establish sufficiency on type spaces where agents have disjoint supports.

It is possible to construct a simple example where EPIC and robust monotonicity are not sufficient for robust monotonicity without lotteries by taking the coordination example of Section 3.3 but removing the outcomes z and z'. As we show in the Appendix, robust implementation is then not possible in this example despite the fact that the social choice function selects a unique strictly Pareto-dominant outcome at every type profile.

6.4 Ex Post and Robust Implementation

In contrast to the earlier results in Bergemann and Morris (2005), where we showed that robust partial implementation is equivalent to ex post incentive compatibility, robust implementation is in general a more demanding notion of implementation than ex post equilibrium implementation. In Bergemann and Morris (2008a) we have analyzed ex post equilibrium implementation. The monotonicity condition there, called ex post monotonicity, is identical to the robust monotonicity condition up to the notion of deception. For ex post monotonicity we have to verify point-to-point deceptions, $\alpha_i : \Theta_i \to \Theta_i$, whereas for robust monotonicity we have to verify point-to-set deceptions $\beta_i : \Theta_i \to 2^{\Theta_i} \big/ \varnothing$. The following simple example, introduced by Palfrey and Srivastava (1989), is useful to relate the different implementation notions and also to understand the role of interdependent types.

Consider a setting with three agents where each agent has two possible "payoff types", θ_a or θ_b. There are two possible choices for society, a or b. All agents have identical preferences. If a majority of agents (i.e., at least two) are of type θ_y, then every agent gets utility 1 from outcome y and utility 0 from the other outcome. The social choice function agrees with the common preferences of the agents. Thus $f : \{\theta_a, \theta_b\}^3 \to \{a, b\}$ satisfies $f(\theta) = y$ if and only if $\#\{i : \theta_i = \theta_y\} \geq 2$.

Clearly, ex post incentive compatibility is not a problem in this example. The problem is that in the "direct mechanism" - where all agents simply announce their types - there is the possibility that all agents will choose to always announce θ_a. Since no agent expects to be pivotal, he has no incentive to truthfully announce his type when he is in fact θ_b. What happens if we allow more complicated mechanisms?

If there were complete information about agents' preferences, then the social choice function is clearly implementable: the social planner could pick an agent, say agent 1, and simply follow that agent's recommendation.

But suppose instead that there is incomplete information about agents' preferences. In particular, suppose it is common knowledge that each agent's type is θ_b with independent probability q, with $q^2 > \frac{1}{2}$. This example fails the Bayesian monotonicity condition of Postlewaite and Schmeidler (1986) and Jackson (1991). Palfrey and Srivastava (1989) observe that it is also not possible to

implement in undominated Bayesian Nash equilibrium in this example.

In contrast, it is easy to construct an augmented mechanism whose only ex post equilibrium delivers the social choice function. Let each agent send a message $m_i \in \{\theta_a, \theta_b\} \times \{\text{truth, lie}\}$, with the interpretation that an agent is announcing his own type and also sends the message "truth" if he thinks that others are telling the truth and sends the message "lie" if he thinks that someone is lying. Outcome y is implemented if a majority claim to be type θ_y and all agents announce "truth"; or if either 1 or 3 agents claim to be type θ_y and at least one agent reports lying.

There is a truthtelling ex post equilibrium where each agent truthfully announces his type and also announces "truth". Now suppose there exists an ex post equilibrium such that at some type profile, the desired outcome is not chosen. Note that whatever the announcements of the other agents, each agent always has the ability to determine the outcome y, by sending the message "lie" and - given the announcements of the other agents - choosing his message so that an odd number of agents have claimed to be type θ_y. So this is not consistent with ex post equilibrium.

Robust implementation is impossible in this example. Consider the type space where there is common knowledge that whenever an agent is type θ_y, he assigns probability $\frac{1}{2}$ to both of the other agents being type $y' \neq y$ and probability $\frac{1}{2}$ to one being type y and the other being y'. Thus every type of every agent thinks there is a 50% chance that outcome a is better and a 50% chance that b is better. Evidently, there is no way of designing a mechanism that ensures that agents do not fully pool. But if they fully pool, robust implementation is not possible.

6.5 Extensions

The previous sections examined the importance of our assumptions about lotteries over outcomes and restrictions on mechanisms. We also restricted attention in our main analysis to the case of discrete but infinite pure outcomes Z, payoff types Θ_i and types T_i. While most of our results would extend naturally to more general Z, Θ_i and T_i, the formal treatment of non-compact type spaces would raise technical issues that we have chosen to avoid.

7 Appendix

7.1 Robust Monotonicity and Dual Robust Monotonicity

Proof of Lemma 2. Suppose that β is (strictly) refutable. Then there exist i, θ_i, $\theta_i' \in \beta_i(\theta_i)$ such that, for all $\theta_{-i}' \in \Theta_{-i}$, there exists a compact set $\overline{Y} \subseteq Y$ such that $y \in \overline{Y}$ implies

$$u_i\left(f\left(\theta_i'', \theta_{-i}\right), \left(\theta_i'', \theta_{-i}'\right)\right) \geq (>)\, u_i\left(y, \left(\theta_i'', \theta_{-i}'\right)\right)$$

for all θ_i'' with $f\left(\theta_i'', \theta_{-i}\right) \neq y$ and, for each $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$, there exists $y \in \overline{Y}$ such that

$$\sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i(\theta_{-i})\, u_i(y, (\theta_i, \theta_{-i})) > \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i(\theta_{-i})\, u_i\left(f\left(\theta_i', \theta_{-i}'\right), (\theta_i, \theta_{-i})\right).$$

By the equivalence between strict domination and never a best response (see Theorem 2.10 in Gale (1989)), there exists $y^* \in \overline{Y}$ with $u_i(y^*, (\theta_i, \theta_{-i})) > u_i\left(f\left(\theta_i', \theta_{-i}'\right), (\theta_i, \theta_{-i})\right)$ for all $\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)$. This establishes the conditions of the lemma.■

Proof of Lemma 3. Suppose conditional NTI holds. Then, for each i, θ_i, θ_{-i}', there exists a compact set $\overline{Y} \triangleq Y_i\left(\theta_{-i}'\right)$ such that $y \in \overline{Y}$ implies

$$u_i\left(f\left(\theta_i'', \theta_{-i}\right), \left(\theta_i'', \theta_{-i}'\right)\right) \geq u_i\left(y, \left(\theta_i'', \theta_{-i}'\right)\right)$$

for all θ_i''. Writing $\overline{y}$ for a lottery strictly in the interior of $\overline{Y}$, we also have that for each $\psi_i \in \Delta(\Theta_{-i})$, there exists $y \in \overline{Y}$ such that

$$\sum_{\theta_{-i} \in \Theta_{-i}} \psi_i(\theta_{-i})\, u_i(y, (\theta_i, \theta_{-i})) > \sum_{\theta_{-i} \in \Theta_{-i}} \psi_i(\theta_{-i})\, u_i(\overline{y}, (\theta_i, \theta_{-i})).$$

By the equivalence between strict domination and never a best response (see Theorem 2.10 in Gale (1989)), there exists $y^* \in \overline{Y}$ with $u_i(y^*, (\theta_i, \theta_{-i})) > u_i(\overline{y}, (\theta_i, \theta_{-i}))$ for all $\theta_{-i} \in \Theta_{-i}$. This establishes the conditions of the lemma. ■

7.2 A Badly Behaved Mechanism

The example illustrates the gap between the necessary and sufficient conditions in Theorem 3. Specifically, it shows that there can be an equilibrium for every type space $\mathcal{T}$ in a mechanism, yet $S^{\mathcal{M}}$ does not satisfy the ex post best response property.

In the example, there are two agents and there is complete information, so each agent has a unique payoff type. There are a finite number of outcomes $Z = \{a, b, c\}$. The payoffs are given by the following table:

	a	b	c
agent 1	0	-1	$+1$
agent 2	0	0	0

.

The planner's choice (in the unique payoff state) is a. Thus it is trivial to robustly implement the social choice function. But suppose that the planner chooses the following (strange) mechanism: $M_1 = \{1, 2, 3, ...\}$, $M_2 = \{1, 2\}$ and

$$g(m_1, m_2) = \begin{cases} a, & \text{if } m_1 = 1; \\ b, & \text{if } m_1 > 1 \text{ and } m_2 = 1; \\ \left[\frac{1}{m_1}, b; \left(1 - \frac{1}{m_1}\right), c\right], & \text{if } m_1 > 1 \text{ and } m_2 = 2. \end{cases}$$

where $\left[\frac{1}{m_1}, b; \left(1 - \frac{1}{m_1}\right), c\right]$ is the lottery putting probability $\frac{1}{m_1}$ on b and probability $\left(1 - \frac{1}{m_1}\right)$ on c. Thus $g(m_1, m_2)$ can be represented by the following table:

g	1	2
1	a	a
2	b	$\left[\frac{1}{2}, b; \frac{1}{2}, c\right]$
3	b	$\left[\frac{1}{3}, b; \frac{2}{3}, c\right]$
$\vdots$	$\vdots$	$\vdots$
k	b	$\left[\frac{1}{k}, b; 1 - \frac{1}{k}, c\right]$
$\vdots$	$\vdots$	$\vdots$

.

Thus the agents are playing the following complete information game:

m_1/m_2	1	2
1	$0, 0$	$0, 0$
2	$-1, 0$	$0, 0$
3	$-1, 0$	$\frac{1}{3}, 0$
$\vdots$	$\vdots$	$\vdots$
k	$-1, 0$	$1 - \frac{2}{k}, 0$
$\vdots$	$\vdots$	$\vdots$

.

Now on any type space, there is always an equilibrium where agent 1 chooses action 1 and agent 2 chooses action 1, and outcome a is chosen. Moreover, on any type space, in any equilibrium, outcome a is always chosen: if agent 1 ever has a best response not to play 1 then he has no best response. So he always plays 1 in equilibrium. Thus the trivial social choice function is robustly implemented by this mechanism.

While only message 1 survives iterated deletion of never best responses for agent 1, both messages survive iterated deletion of never best responses for agent 2. Thus we have $S_1^{\mathcal{M}} = \{1\}$ and $S_2^{\mathcal{M}} = \{1, 2\}$. Note that $S^{\mathcal{M}}$ does not satisfy the interim best response property as we observe that

$$u_1(g(1, 2)) = u_1(a) = 0 < \frac{1}{3} = u_1(g(3, 2)),$$

violating the ex post best response property.

The insight of the example is that the quantifier "for every type space $\mathcal{T}$" does not necessarily guarantee that all actions which will be chosen with positive probability in some equilibrium and for some type space, will also be chosen with probability one in some equilibrium for some type space. For this reason, the quantifier "for every type space $\mathcal{T}$" does not allow us to establish a local, i.e. ex post best response property of every action in $S^{\mathcal{M}}$.

7.3 Coordination Example Continued

The final example is the pure coordination game, which we first considered in Section 3.3, but without the additional allocations, z and z'. It illustrates the importance of lotteries for robust implementation. The example will satisfy EPIC and robust monotonicity, yet it cannot be robustly implemented without the use of lotteries. On the other hand the preferences clearly satisfy the conditional NTI property, and hence the sufficient conditions for robust implementation would be satisfied with lotteries.

As in the example in Section 3.3, the payoffs of the player are given by (11) and the social choice function f is given by (12). The social choice function is strictly ex post incentive compatible but there is another equilibrium in the "direct mechanism" where each agent misreports his type, and each agent gets a payoff of 1.

Robust monotonicity is clearly satisfied even if the rewards $Y_i(\theta_{-i})$ are restricted to the deterministic allocations Z. We next show that robust implementation is not possible even in an infinite mechanism if we restrict attention to deterministic mechanisms. Fix a mechanism $\mathcal{M}$. Let

$$S_i^*(\theta_i) = \{m_i : g(m_i, m_j) = f(\theta_i, \theta_j) \text{ for some } m_j, \theta_j\},$$

be the set of messages for agent i which would select the allocation recommended by the social choice function for some m_j, θ_j. We establish by induction that, $S_i^*(\theta_i) \subseteq S_i^k(\theta_i)$ for all k using the structure of the payoffs. Suppose that this is true for k. Then for any $m_i \in S_i^*(\theta_i) \subseteq S_i^k(\theta_i)$, there exists $m_j \in S_j^*(\theta_j) \subseteq S_j^k(\theta_j)$ such that $g(m_i, m_j) = f(\theta_i, \theta_j)$. Thus there does not exist $\nu_i \in \Delta(M_i)$ such that

$$\sum_{m_i'} \nu_i(m_i') u_i\left(g\left(m_i', m_j\right), (\theta_i, \theta_j)\right) > u_i(g(m_i, m_j), (\theta_i, \theta_j)) = 3,$$

and so $m_i \in S_i^{k+1}(\theta_i)$.

Thus we must have that $(m_1, m_2) \in S_1^*(\theta_1) \times S_2^*(\theta_2)$ implies $g(m_1, m_2) = f(\theta_1, \theta_2)$. Let $m_i^*(\cdot)$ be any selection from $S_i^*(\cdot)$. Now let k^* be the lowest k such that, for some i,

$$m_i^*(\theta_i') \notin S_i^k(\theta_i).$$

Without loss of generality, let $i = 1$. Note $m_2^* \left(\theta_2'\right) \in S_2^k \left(\theta_2\right)$ for all $k < k^*$ by definition of k^*. If agent 1 was type θ_1 and was sure his opponent were type θ_2 and choosing action $m_2^* \left(\theta_2'\right)$, we know that he could guarantee himself a payoff of 1 by choosing $m_1^* \left(\theta_1'\right)$. Since $m_1^* \left(\theta_1'\right)$ is deleted for type θ_1 at round k^*, we know that there exists $\nu_1 \in \Delta \left(M_1\right)$ such that:

$$\sum_{m_1'} \nu_1 \left(m_1'\right) g_1 \left(m_1', m_2^* \left(\theta_2'\right)\right) > 1,$$

and thus there exists m_1' such that $g_1 \left(m_1', m_2^* \left(\theta_2'\right)\right) = f \left(\theta_1, \theta_2\right)$. This implies that $m_2^* \left(\theta_2'\right) \in S_2^* \left(\theta_2\right)$, a contradiction.

The example uses the fact that the social choice function always selects an outcome that is strictly Pareto-optimal and - paradoxically - it is this feature which inhibits rationalizable implementation in the current example. Börgers (1995) proves the impossibility of complete information implementation of non-dictatorial social choice functions in iteratively undominated strategies when the set of feasible preference profiles includes such unanimous preference profiles and the argument here is reminiscent of Börgers' argument.

7.4 Responsive Social Choice Functions

Here we provide the formal statements and proofs for the discussion in Subsection 6.1.

Lemma 4 *If social choice function is responsive and satisfies semi-strict EPIC, then it satisfies strict EPIC.*

Proof. This follows directly from definitions. ■

Lemma 5 *If f satisfies strict EPIC and strict pairwise robust monotonicity, then f satisfies strict robust monotonicity.*

Proof. By strict pairwise robust monotonicity if, for every unacceptable deception β, there exist i, θ_i and θ_i' such that, for all $\theta_{-i}' \in \Theta_{-i}$ and $\psi_i \in \Delta \left(\beta_{-i}^{-1} \left(\theta_{-i}'\right)\right)$, there exists $y \in \Delta \left(Z\right)$ such that

$$\sum_{\theta_{-i}} \psi_i \left(\theta_{-i}\right) u_i \left(y, \theta\right) > \sum_{\theta_{-i}} \psi_i \left(\theta_{-i}\right) u_i \left(f \left(\theta'\right), \theta\right), \tag{19}$$

and

$$u_i \left(f \left(\theta_i', \theta_{-i}'\right), \left(\theta_i', \theta_{-i}'\right)\right) > u_i \left(y, \left(\theta_i', \theta_{-i}'\right)\right). \tag{20}$$

Now fix $\theta_i'' \neq \theta_i$. Strict EPIC implies

$$u_i \left(f \left(\theta_i'', \theta_{-i}'\right), \left(\theta_i'', \theta_{-i}'\right)\right) > u_i \left(f \left(\theta_i', \theta_{-i}'\right), \left(\theta_i'', \theta_{-i}'\right)\right). \tag{21}$$

Now we can define a lottery y^{ε} by:

$$y^{\varepsilon} \triangleq (1-\varepsilon) f\left(\theta_i', \theta_{-i}'\right) + \varepsilon y.$$

Now (19) and (20) continue to hold, replacing y with y^{ε}. For sufficiently small $\varepsilon > 0$, we also have

$$u_i\left(f\left(\theta_i'', \theta_{-i}'\right), \left(\theta_i'', \theta_{-i}'\right)\right) > u_i\left(y^{\varepsilon}, \left(\theta_i'', \theta_{-i}'\right)\right)$$

for all θ_i''. Thus strict robust monotonicity is satisfied. ■

Now consider the following weakening of rationalizable implementation which weakens part (c) of Definition 4.2. In particular, it drops the earlier uniformity condition which required that a constant belief $\lambda_i \in \Delta\left(M_{-i} \times \Theta_{-i}\right)$ would support the existence of a best response for all payoff types $\theta_i \in \Theta_i$. In the current weaker version, the belief $\lambda_i \in \Delta\left(M_{-i} \times \Theta_{-i}\right)$ is allowed to depend on θ_i.

Definition 17 (Weak Rationalizable Implementation)
Social choice function f is weakly rationalizably implementable by mechanism $\mathcal{M}$ if:

1. $m \in S^{\mathcal{M}}(\theta) \Rightarrow g(m) = f(\theta)$; *and*

2. *for all i, θ_i and $\psi_i \in \Delta\left(\Theta_{-i}\right)$, there exists $\lambda_i \in \Delta\left(M_{-i} \times \Theta_{-i}\right)$ such that:*

$$\begin{aligned}
&(a)\ \lambda_i\left(m_{-i}, \theta_{-i}\right) > 0 \Rightarrow m_j \in S_j^{\mathcal{M}}\left(\theta_j\right) \text{ for all } j \neq i, \\
&(b)\ \sum_{m_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) = \psi_i\left(\theta_{-i}\right) \text{ for all } \theta_{-i} \in \Theta_{-i}, \\
&(c)\ \arg\max_{m_i} \sum_{m_{-i}, \theta_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) u_i\left(g\left(m_i, m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right) \neq \varnothing.
\end{aligned}$$

Social choice function f is weakly rationalizably implementable if there exists a mechanism $\mathcal{M}$ such that f is weakly rationalizably implementable by $\mathcal{M}$.

We next establish the necessary conditions under weak rationalizability.

Lemma 6 *If f is weakly rationalizably implementable, then f satisfies semi-strict EPIC.*

Proof. Let $\mathcal{M}$ be a mechanism such that f is weakly rationalizably implementable by $\mathcal{M}$. We write $\psi_i^{\theta_{-i}}$ for the probability distribution putting probability one on $\theta_{-i} \in \Theta_{-i}$. Now for each $\theta \in \Theta$, the definition of weak rationalizable implementation (17) requires that there exists $\nu_i^{\theta} \in \Delta\left(S_{-i}^{\mathcal{M}}\left(\theta_{-i}\right)\right)$ such that

$$\arg\max_{m_i} \sum_{m_{-i}} \nu_i^{\theta}\left(m_{-i}\right) u_i\left(g\left(m_i, m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right) \neq \varnothing; \tag{22}$$

where $\nu_i^\theta \in \Delta\left(S_{-i}^{\mathcal{M}}(\theta_{-i})\right)$ is a probability distribution over the rationalizable messages at the type profile θ_{-i}. But the definition of rationalizable implementation (17) also requires that $m \in S^{\mathcal{M}}(\theta) \Rightarrow g(m) = f(\theta)$ and thus

$$\sum_{m_{-i}} \nu_i^\theta(m_{-i})\, u_i\left(g(m_i, m_{-i}), (\theta_i, \theta_{-i})\right) = u_i\left(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\right) \tag{23}$$

for all $m_i \in S_i^{\mathcal{M}}(\theta_i)$. Now (22) and (23) imply

$$u_i\left(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\right) \geq \sum_{m_{-i}} \nu_i^\theta(m_{-i})\, u_i\left(g(m_i, m_{-i}), (\theta_i, \theta_{-i})\right), \tag{24}$$

for all $m_i \in M_i$. Now (24) implies that, for all $m_i' \in S_i^{\mathcal{M}}\left(\theta_i'\right)$:

$$\begin{aligned} u_i\left(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\right) &\geq \sum_{m_{-i}} \nu_i^\theta(m_{-i})\, u_i\left(g\left(m_i', m_{-i}\right), (\theta_i, \theta_{-i})\right) \\ &= u_i\left(f\left(\theta_i', \theta_{-i}\right), (\theta_i, \theta_{-i})\right), \text{ for all } \theta_i' \in \Theta_i, \end{aligned} \tag{25}$$

which establishes semi-strict EPIC. ■

We also use the contra-positive statements of strict pairwise robust monotonicity in terms of pairwise refutability.

Definition 18 (Strict Pairwise Refutable)
Deception β is strictly pairwise refutable if there exist i, θ_i and $\theta_i' \in \beta_i(\theta_i)$ such that, for all $\theta_{-i}' \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$, there exists $y \in \Delta(Z)$ such that

$$\sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i(y, \theta) > \sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i\left(f\left(\theta'\right), \theta\right),$$

and

$$u_i\left(f\left(\theta_i', \theta_{-i}'\right), \left(\theta_i', \theta_{-i}'\right)\right) > u_i\left(y, \left(\theta_i', \theta_{-i}'\right)\right).$$

Conversely, we define when a deception β is not strictly refutable.

Definition 19 (Not Strict Pairwise Refutable)
Deception β is not strictly pairwise refutable if for all i, θ_i and $\theta_i' \in \beta_i(\theta_i)$, there exist $\theta_{-i}' \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$ such that:

$$u_i\left(y, \left(\theta_i', \theta_{-i}'\right)\right) \geq u_i\left(f\left(\theta_i', \theta_{-i}'\right), \left(\theta_i', \theta_{-i}'\right)\right)$$

implies that

$$\sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i(y, \theta) > \sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i\left(f\left(\theta'\right), \theta\right).$$

We can then restate pairwise robust monotonicity in terms of refutable deception. In particular, a social choice function f satisfies strict pairwise robust monotonicity if and only if every strictly pairwise refutable deception is unacceptable. This allows to obtain necessary conditions for weak rationalizable implementation if the social choice function f is responsive.

Proposition 4 *If f is responsive and weakly rationalizably implementable, then f satisfies strict pairwise robust monotonicity.*

Proof. Suppose that f is weakly rationalizable implementable by $\mathcal{M}$. Suppose that β is not strictly pairwise refutable. Define S^{β} by

$$S_i^{\beta}(\theta_i) \triangleq \bigcup_{\theta_i' \in \beta_i(\theta_i)} S_i^{\mathcal{M}}(\theta_i').$$

We claim

$$S^{\beta} \leq b\left(S^{\beta}\right).$$

To see why, fix any i, θ_i, $\theta_i' \in \beta_i(\theta_i)$. Because β is not strictly refutable, there exist $\theta_{-i}' \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$ such that

$$u_i\left(f\left(\theta_i', \theta_{-i}'\right), \left(\theta_i, \theta_{-i}'\right)\right) \geq u_i\left(y, \left(\theta_i', \theta_{-i}'\right)\right)$$

implies

$$\sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i(y, \theta) > \sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i\left(f\left(\theta'\right), \theta\right).$$

Part (2) of the definition of weak rationalizable implementation implies that there exists $\nu_i^{\theta_{-i}'} \in \Delta\left(S_{-i}^{\mathcal{M}}\left(\theta_{-i}'\right)\right)$ such that

$$\arg\max_{m_i} \sum_{m_{-i}} \nu_i^{\theta_{-i}'}(m_{-i})\, u_i\left(g(m_i, m_{-i}), \left(\theta_i', \theta_{-i}'\right)\right) \neq \varnothing.$$

Now, for every $\widetilde{m}_i \in M_i$, define

$$y(\widetilde{m}_i) \triangleq \sum_{m_{-i}} \nu_i^{\theta_{-i}'}(m_{-i})\, g(\widetilde{m}_i, m_{-i}).$$

Now if $y(\widetilde{m}_i) \neq f\left(\theta_i', \theta_{-i}'\right)$, then we must have

$$u_i\left(f\left(\theta_i', \theta_{-i}'\right), \left(\theta_i', \theta_{-i}'\right)\right) \geq u_i\left(y(\widetilde{m}_i), \left(\theta_i', \theta_{-i}'\right)\right)$$

(if not, $\widetilde{m}_i$ would be rationalizable for type θ_i', contradicting rationalizable implementation). But now, by (7),

$$\sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i(\theta_{-i})\, u_i\left(f\left(\theta_i', \theta_{-i}'\right), (\theta_i, \theta_{-i})\right) \geq \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i(\theta_{-i})\, u_i\left(y(\widetilde{m}_i), (\theta_i, \theta_{-i})\right)$$

for all $\widetilde{m}_i$, and thus, for any $m_i' \in S_i^{\mathcal{M}}\left(\theta_i'\right)$ we have:

$$\begin{aligned}
& \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right), m_{-i}} \psi_i\left(\theta_{-i}\right) \nu_i^{\theta_{-i}'}\left(m_{-i}\right) u_i\left(g\left(m_i', m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right) \\
= & \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i\left(\theta_{-i}\right) u_i\left(f\left(\theta_i', \theta_{-i}'\right), \left(\theta_i, \theta_{-i}\right)\right) \\
\geq & \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i\left(\theta_{-i}\right) u_i\left(y\left(\widetilde{m}_i\right), \left(\theta_i, \theta_{-i}\right)\right) \\
= & \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i\left(\theta_{-i}\right) \nu_i^{\theta_{-i}'}\left(m_{-i}\right) u_i\left(g\left(\widetilde{m}_i, m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right).
\end{aligned}$$

But now $S^\beta \leq b\left(S^\beta\right) \Rightarrow S^\beta \leq S^{\mathcal{M}} \Rightarrow \beta$ is acceptable. ■

References

Abreu, D., and H. Matsushima (1992a): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993–1008.

——— (1992b): "Virtual Implementation In Iteratively Undominated Strategies: Incomplete Information," Discussion paper, Princeton University and University of Tokyo.

Battigalli, P., and M. Siniscalchi (2003): "Rationalization and Incomplete Information," *Advances in Theoretical Economics*, 3, Article 3.

Benoit, J.-P., and E. Ok (2008): "Nash Implementation Without No-Veto Power," *Games and Economic Behavior*, 64, 51–67.

Bergemann, D., and S. Morris (2005): "Robust Mechanism Design," *Econometrica*, 73, 1771–1813.

——— (2008a): "Ex Post Implementation," *Games and Economic Behavior*, 63, 527–566.

——— (2008b): "Robust Implementation in General Mechanisms," Discussion Paper 1666, Cowles Foundation, Yale University.

——— (2009a): "Robust Implementation in Direct Mechanisms," *Review of Economic Studies*, 76, 1175–1206.

——— (2009b): "Robust Virtual Implementation," *Theoretical Economics*, 4, 45–88.

Bergemann, D., S. Morris, and O. Tercieux (2010): "Rationalizable Implementation," Discussion Paper CFDP 1697R, Cowles Foundation, Yale University.

Bochet, O. (2007): "Nash Implementation with Lottery Mechanisms," *Social Choice and Welfare*, 28, 111–125.

Börgers, T. (1995): "A Note on Implementation and Strong Dominance," in *Social Choice, Welfare and Ethics*, ed. by W. Barnett, H. Moulin, M. Salles, and N. Schofield. Cambridge University Press, Cambridge.

Brandenburger, A., and E. Dekel (1987): "Rationalizability and Correlated Equilibria," *Econometrica*, 55, 1391–1402.

Chung, K.-S., and J. Ely (2001): "Efficient and Dominance Solvable Auctions with Interdependent Valuations," Discussion paper, Northwestern University.

Dekel, E., D. Fudenberg, and S. Morris (2007): "Interim Correlated Rationalizability," *Theoretical Economics*, 2, 15–40.

Duggan, J. (1997): "Virtual Bayesian Implementation," *Econometrica*, 65, 1175–1199.

Gale, D. (1989): *The Theory of Linear Economic Models*. University of Chicago Press, Chicago.

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

Jackson, M. (1992): "Implementation in Undominated Strategies: A Look at Bounded Mechanisms," *Review of Economic Studies*, 59, 757–775.

Lipman, B. (1994): "A Note on the Implications of Common Knowledge of Rationality," *Games and Economic Behavior*, 6, 114–129.

Palfrey, T., and S. Srivastava (1989): "Mechanism Design with Incomplete Information: A Solution to the Implementation Problem," *Journal of Political Economy*, 97, 668–691.

Postlewaite, A., and D. Schmeidler (1986): "Implementation in Differential Information Economies," *Journal of Economic Theory*, 39, 14–33.

Serrano, R., and R. Vohra (2005): "A Characterization of Virtual Bayesian Implementation," *Games and Economic Behavior*, 50, 312–331.

Wilson, R. (1987): "Game-Theoretic Analyses of Trading Processes," in *Advances in Economic Theory: Fifth World Congress*, ed. by T. Bewley, pp. 33–70, Cambridge. Cambridge University Press.