Theoretical Economics 4 (2009), 45–88 1555-7561/20090045

Robust virtual implementation

Dirk Bergemann
Department of Economics, Yale University

Stephen Morris
Department of Economics, Princeton University

In a general interdependent preference environment, we characterize when two payoff types can be distinguished by their rationalizable strategic choices without any prior knowledge of their beliefs and higher order beliefs. We show that two payoff types are strategically distinguishable if and only if they satisfy a separability condition. The separability condition for each agent essentially requires that there is not too much interdependence in preferences across agents.

A social choice function, mapping payoff type profiles to outcomes, can be robustly virtually implemented if there exists a mechanism such that every equilibrium on every type space achieves an outcome that is arbitrarily close to the outcome generated by the social choice function. This definition is equivalent to requiring virtual implementation in iterated deletion of strategies that are strictly dominated for all beliefs. The social choice function is robustly measurable if strategically indistinguishable payoff types receive the same allocation. We show that ex post incentive compatibility and robust measurability are necessary and sufficient for robust virtual implementation.

Keywords. Mechanism design, virtual implementation, robust implementation, rationalizability, ex post incentive compatibility.

JEL classification. C79, D82.

1. Introduction

Suppose that a social planner would like to design a mechanism that induces self-interested agents to make strategic choices that lead to the selection of socially desirable outcomes. A *social choice function* specifies the socially desired outcomes as a function of the unobserved payoff types of the agents. The planner would like to be sure that outcomes specified by the social choice function arise with probability arbitrarily close

Dirk Bergemann: dirk.bergemann@yale.edu
Stephen Morris: smorris@princeton.edu

This research is partially supported by NSF Grants #CNS 0428422 and #SES-0518929. We thank the co-editor, Jeffrey Ely, and two anonymous referees for their valuable comments and suggestions. We are grateful for discussions with Dilip Abreu, Faruk Gul, Matt Jackson, Eric Maskin, Wolfgang Pesendorfer, Phil Reny, Roberto Serrano, and seminar audiences at Chicago, Georgetown, Hong Kong, IMPA, NUS, Penn, Princeton, Rutgers, and UT Austin. This paper incorporates and replaces some preliminary results on robust virtual implementation appearing in Bergemann and Morris (2005a).

to 1: thus she requires *virtual* implementation. In addition, she would like *every* possible equilibrium to virtually implement the social choice function: thus she requires *full* implementation. Finally she would like every equilibrium to virtually implement the social choice function whatever the agents' beliefs and higher order beliefs about others' types; thus she requires *robust* implementation. In this paper, we provide a characterization of when *robust virtual implementation* is possible in a general interdependent preference environment.

One necessary condition for robust virtual implementation is *ex post incentive compatibility*: under the social choice function, each agent must have an incentive to truthfully report his type if others report their types truthfully, whatever their types. Ex post incentive compatibility is sufficient to ensure the existence of desirable equilibria, but, as the existing incomplete information implementation literature emphasizes, further restrictions on the social choice function are required to rule out other, undesirable, equilibria. If a mechanism is to fully implement a social choice function, two types who are treated differently by the social choice function must be guaranteed to behave differently in the implementing mechanism. The key result in this paper is a characterization of when two payoff types are *strategically distinguishable* in this sense that they can be guaranteed to behave differently. A second necessary condition for robust virtual implementation is *robust measurability*: strategically indistinguishable types are treated the same by the social choice function. We show that ex post incentive compatibility and robust measurability are also sufficient for robust virtual implementation (under an economic assumption).

Thus the core of our contribution is an analysis of strategic distinguishability. Fix an interdependent preferences environment, with a finite set of agents, each with a finite set of possible *payoff types*, with expected utility preferences over lotteries depending on the whole profile of types. Two payoff types of an agent are *strategically distinguishable* if they have disjoint rationalizable strategic choices in some finite game for all possible beliefs and higher order beliefs about others' types. Thus two payoff types are strategically *indistinguishable* if in every game, there exists some action that each type might rationally choose given some beliefs and higher order beliefs. We are able to provide an exact and insightful characterization of strategic distinguishability. If we have sets of types, Ψ_1 and Ψ_2, of agents 1 and 2, respectively, we say that Ψ_2 separates Ψ_1 if knowing agent 1's preferences and knowing that agent 1 is sure that agent 2's type is in Ψ_2, we can rule out at least one type of agent 1. Now consider an iterative process where we start, for each agent, with all subsets of his type set and, at each stage, delete subsets of actions that are separated by every remaining subset of types of his opponents. A pair of types is said to be *pairwise inseparable* if the set consisting of that pair of types survives this process. We show that two types are strategically indistinguishable if and only if they are pairwise inseparable.

If there are private values and every type is value distinguished, then every pair of types is pairwise separable and thus strategically distinguishable. Thus strategic indistinguishability arises when the degree of interdependence in preferences is large. We can illustrate this with a simple example. Suppose that agent i's payoff type is $\theta_i \in [0,1]$

and agent i's valuation of a private good is $\theta_i + \gamma \sum_{j \neq i} \theta_j$. Each agent has quasilinear utility, i.e., his utility from money is linear and additive. We show that all distinct pairs of types are strategically distinguishable if $|\gamma| < 1/(I-1)$, where I is the number of agents. All pairs of types are strategically indistinguishable if $|\gamma| \geq 1/(I-1)$.

Our characterization result for strategic distinguishability (Theorem 1) comes in two parts. If two types of an agent are pairwise inseparable, then they belong to a set of types that are not separable by a profile of sets of types of that agent's opponents. The set of types of each opponent in that profile is then not separable by a profile of sets of types of that opponent's opponents. And there is a continuing chain of inseparable sets in the chain. We prove that pairwise inseparable types are strategically indistinguishable (Proposition 1) by induction, showing that in any mechanism at any stage in the iterated deletion of messages that are never best responses and for every set of types in the chain of inseparable type sets, a common action is played. The inseparability property ensures that we can always construct beliefs for each type that make the same message a best response.

To show the converse result (Proposition 2), we construct a finite *maximally revealing mechanism* with the property that all pairwise separable types have disjoint sets of rationalizable actions. The construction exploits the linearity of expected utility preferences and duality theory. Whenever a set of types of one agent is separated by a profile of sets of types of other agents, we are able to construct a finite set of lotteries such that knowing the first agent's preference over those lotteries always rules out at least one of his types. We can take the union over all such finite sets constructed for each profile of type sets where the separability property holds. We then construct a finite "test set" of lotteries such that knowing an agent's most preferred outcome in that test set implicitly reveals his ranking of outcomes in all the original sets. Finally, we consider a mechanism where each agent gets to pick a lottery with some positive probability, then guesses which lotteries others chose and gets to pick another lottery, with small probability, contingent on other agents making the choice he conjectured, and so on. With a large, but finite, number of stages this mechanism eventually leads pairwise separable types to make distinct choices.

Our proof of the sufficiency of ex post incentive compatibility and robust measurability (Corollary 1) for robust virtual implementation builds on an ingenious construction used by Abreu and Matsushima (1992b) to establish an extremely permissive result for complete information virtual implementation. In Abreu and Matsushima (1992c), they adapt the argument to a standard Bayesian virtual implementation problem; we in turn adapt the argument to our robust virtual implementation problem.

While our sufficiency argument for robust virtual implementation builds on Abreu and Matsushima (1992c), the interpretation of our results ends up being rather different. Abreu and Matsushima (1992c) characterize virtual implementation in a standard Bayesian environment, where there is common knowledge of a common prior over a fixed set of types, using the solution concept of iterated deletion of strictly dominated strategies and restricting attention to well-behaved (finite) mechanisms. Bayesian incentive compatibility of the social choice function is a necessary condition: a standard

compactness argument shows that the weakening to *virtual* implementation does not weaken the incentive compatibility requirement. In addition, they show that a *measurability* condition is necessary. Put each agent's types into equivalence classes that have the same preferences over outcomes, unconditional on other agents' types. Having distinguished some types by their unconditional preferences, we can then further refine agents' types, by distinguishing types with different preferences conditional on other agents' types in the first stage. We can continue this process of refining agents' types based on preferences conditional on other agents' types revealed so far. The social choice function is *Abreu–Matsushima measurable* if it is measurable with respect to the limit of this iterative refinement. This seems to be a weak restriction that is generically satisfied.[1] Abreu and Matsushima (1992c) show that Bayesian incentive compatibility and Abreu–Matsushima measurability are sufficient as well as necessary for virtual implementation in iterated deletion of strictly dominated strategies.

Robust virtual implementation is equivalent to requiring that there is a single mechanism that implements a social choice function, for all possible type spaces that could be constructed for the environment with fixed payoff types and utility functions for the agents. It is instructive to see how to get from Abreu and Matsushima (1992c) to the robust virtual implementation results in this paper.

Observe that Abreu and Matsushima (1992c)'s solution concept naturally uses agents' given beliefs about others' types: when strategies are deleted, it is because they are strictly dominated conditional on the agents' beliefs. We want implementation for all possible beliefs. We therefore establish our results under an incomplete information version of rationalizability that does not make use of any beliefs over others' types; it is equivalent to iteratively deleting strategies that are *ex post strictly dominated*, i.e., strictly dominated for all possible beliefs over others' types. We work with this solution concept throughout the paper. However, results from the epistemic foundations of game theory establish that an action is rationalizable in this sense for a payoff type if and only if it could be played in an equilibrium on some type space with beliefs and higher order beliefs, by a type with that payoff type (Brandenburger and Dekel 1987, Battigalli and Siniscalchi 2003, and Bergemann and Morris 2008). Thus a bonus of our "robust" analysis is that the distinction between equilibrium and rationalizability (or iterated deletion of strictly dominated strategies) becomes moot.

Now, ex post incentive compatibility is the robust analogue of Bayesian incentive compatibility and robust measurability is the robust analogue of the measurability of Abreu and Matsushima (1992c). They can reasonably argue that, in a standard Bayesian setting, their measurability condition is a weak technical requirement.[2] As a result, the "bottom line" of the virtual implementation literature has been that full implementation, i.e., getting rid of undesirable equilibria, does not impose any substantive

[1] Abreu and Matsushima (1992c) and Serrano and Vohra (2005) note that a simple sufficient condition for all social choice functions to be A–M measurable is *type diversity*: every type has distinct preferences over lotteries unconditional on others' types.

[2] Although Serrano and Vohra (2001) describe an economic example where all non-trivial individually rational and Bayesian incentive compatible social choice functions fail Abreu–Matsushima measurability because types have identical conditional preferences.

constraints beyond incentive compatibility, i.e., the existence of desirable equilibria. By requiring the more demanding, but more plausible, robust formulation of incomplete information, we end up with a condition that is substantive (imposing significantly more structure in interdependent value environments than incentive compatibility) and easily interpretable.

This paper adds to a recent literature on robust mechanism design that provides one operationalization of the so-called "Wilson doctrine" that progress in practical mechanism design will come from relaxing the implicit common knowledge assumption in the formulation of mechanism design problems.[3] Neeman (2004) highlights the fact that full surplus extraction with correlated type results (Myerson 1981 and Crémer and McLean 1985) rely on the implicit assumption that there is common knowledge of a mapping from beliefs to payoff types of all agents (a "beliefs determine preferences" property). This (counterintuitive) assumption is implied by the "generic" choice of a common prior on a fixed type space where distinct types are assumed to have different preferences. The apparent weakness of the Abreu–Matsushima measurability condition (and the fact that it is satisfied for "generic" priors) relies on the same property. We believe that by relaxing this unnatural implicit assumption, we get a better insight into the nature of the extra requirement for full implementation over and above incentive compatibility conditions.

Our operationalization of the "Wilson doctrine" is rather strong: we put no restrictions on agents' beliefs and higher order beliefs. A recent paper of Artemov et al. (2008) examines what happens to the conditions for robust virtual implementation if the planner is given partial information about agents' beliefs, in particular, a subset of beliefs over others' payoffs types that can arise with each payoff type. We discuss this intermediate robustness approach in Section 6.3.

It is possible to interpret our result as rather negative: ex post incentive compatibility is already a very strong condition, as emphasized by the recent work of Jehiel et al. (2006).[4] Robust measurability adds the further substantive restriction that there not be too much interdependence of preferences; and, in any case, the mechanism that we use to robustly virtually implement social choice functions is complicated to describe and presumably hard to play. However, we can show that in one large and interesting class of economic environments with interdependent preferences, robust virtual implementation is not only possible but is possible in the direct mechanism where agents simply report their payoff types. Say that an environment has *aggregator single crossing* preferences if the profile of agents' types can be aggregated into a single number and preferences are single crossing with respect to that number. Efficient social choice functions satisfying ex post incentive compatibility often exist in such environments. Bergemann and Morris (forthcoming) show that in such an environment, exact robust implementation is possible if the social choice function satisfies strict ex post incentive compatibility and a contraction property. In this paper, we observe that the contraction

[3] Neeman (2004), Bergemann and Morris (2005b), Heifetz and Neeman (2006), and Chung and Ely (2007).

[4] Although we argue in Bergemann and Morris (forthcoming) that ex post incentive compatibility is feasible in many economically important environments either because types are one-dimensional or because natural economic features of the environment lead to a failure of the "generic" properties that lead to the non-existence of non-trivial ex post incentive compatible social choice functions in Jehiel et al. (2006).

property is equivalent to robust measurability, so that, under the weak condition that there exists some strictly ex post incentive compatible social choice function, whenever robust virtual implementation is possible, it is possible in the direct mechanism.

The remainder of the paper is organized as follows. Section 2 introduces the environment and the solution concept. Section 3 illustrates the notion of separability in the context of a single private good with interdependent preferences. Section 4 defines and characterizes strategic distinguishability, constructing the maximally revealing mechanism to show the equivalence between strategic distinguishability and pairwise separability. Section 5 reports our results on robust virtual implementation. Section 6 concludes with discussions of the formal relation between Abreu–Matsushima measurability and robust measurability, the role of moderate interdependence, intermediate notions of robustness, the epistemic foundations for the solution concept, weak rather than strict dominance, positive results in direct mechanisms, and the relation to exact rather than virtual implementation.

2. Setting

2.1 *Environment*

There is a finite set of agents $\{1,\dots,I\}$ and each agent i has a finite set of possible payoff types

$$\Theta_i = \{\theta_i^1,\dots,\theta_i^s,\dots,\theta_i^S\}.$$

We assume without loss of generality that the cardinality of each set Θ_i is equal to S for all i. The finite set X of pure outcomes is given by

$$X = \{x_1,\dots,x_n,\dots,x_N\}.$$

The lottery space over the set of outcome is $Y = \Delta(X)$. A lottery y is an N-dimensional vector $y = (y_1,\dots,y_n,\dots,y_N)$ with

$$y_n \geq 0 \quad \text{and} \quad \sum_{n=1}^{N} y_n = 1.$$

Each agent has a von Neumann–Morgenstern expected utility function $u_i : Y \times \Theta \to \mathbb{R}$ with

$$u_i(y,\theta) = \sum_{n=1}^{N} u_i(x_n,\theta) y_n.$$

We abuse notation by writing x for the lottery putting probability 1 on outcome x and X for the set of degenerate lotteries.

It is often convenient to work with underlying preferences over lotteries rather than any of their representations. We write $\mathscr{R}$ for the collection of expected utility preference relations on Y. We write $R_{\theta_i,\lambda_i} \in \mathscr{R}$ for the preference relation of agent i if his payoff type is θ_i and he has belief $\lambda_i \in \Delta(\Theta_{-i})$ about the types of others:

$$\forall y, y' \in Y \quad y \, R_{\theta_i,\lambda_i} \, y' \Leftrightarrow \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i}) u_i(y,(\theta_i,\theta_{-i})) \geq \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i}) u_i(y',(\theta_i,\theta_{-i}))$$

and we write P_{θ_i,λ_i} for the strict preference relation corresponding to R_{θ_i,λ_i}.

We make a weak assumption on the preferences: every agent i, whatever his type $\theta_i \in \Theta_i$ and beliefs $\lambda_i \in \Delta(\Theta_{-i})$, has a strict preference over some pair of outcomes.

ASSUMPTION 1 (No Complete Indifference). For each i, $\theta_i \in \Theta_i$, and $\lambda_i \in \Delta(\Theta_{-i})$, there exist $x, x' \in X$ such that $x \, P_{\theta_i,\lambda_i} \, x'$.

We maintain this assumption throughout the paper.[5] An analogous condition appears in Abreu and Matsushima (1992c) and Serrano and Vohra (2005) in the Bayesian setting for all types (and associated beliefs) of all agents. But in our robust context, it is a stronger assumption in the sense that it rules out the possibility that alternative payoff type profiles of others lead to a reversal in the preferences of agent i with respect to some x and x'.

We denote by $\bar{y}$ the *central lottery* that puts equal probability on each of the pure outcomes. Now no-complete-indifference implies that every agent i, whatever his type θ_i and beliefs $\lambda_i \in \Delta(\Theta_{-i})$, strictly prefers some pure outcome x to $\bar{y}$; and compactness implies that those strict preferences are uniformly strict.

LEMMA 1. *There exists $c > 0$ such that, for each i, $\theta_i \in \Theta_i$, and $\lambda_i \in \Delta(\Theta_{-i})$, there exists $x \in X$ such that*

$$\sum_{\theta_{-i} \in \Theta_{-i}} \lambda_i(\theta_{-i}) u_i(x, (\theta_i, \theta_{-i})) > \sum_{\theta_{-i} \in \Theta_{-i}} \lambda_i(\theta_{-i}) u_i(\bar{y}, (\theta_i, \theta_{-i})) + c.$$

This lemma is proved in the Appendix. We use c in our later constructions. We also exploit the existence of an upper bound on payoff differences C that follows immediately from the finiteness of pure outcomes and states

LEMMA 2. *There exists $C > 0$ such that*

$$|u_i(y, \theta) - u_i(y', \theta)| \leq C$$

for all i, y, y', and θ.

2.2 *Mechanisms and solution concept*

A mechanism $\mathcal{M}$ is a collection $((M_i)_{i=1}^I, g)$ where each M_i is finite and $g : M \to Y$. We denote a belief of agent i over the product of payoff type and message spaces of the other agents by $\mu_i \in \Delta(\Theta_{-i} \times M_{-i})$. We consider the process of iteratively eliminating never best responses, without making assumptions on agents' beliefs about others' payoff types. The set of messages surviving the kth level of elimination for type θ_i in mechanism $\mathcal{M}$ is defined by

$$S_i^0[\mathcal{M}](\theta_i) \triangleq M_i$$

[5] Our results can be extended to allow for the presence of complete indifference as shown in the appendix of the working paper version, Bergemann and Morris (2007).

and for each $k=0,\ldots$ by induction:

$$S_i^{k+1}[\mathscr{M}](\theta_i)$$
$$\triangleq \left\{ m_i \in S_i^k[\mathscr{M}](\theta_i) \;\middle|\; \begin{array}{l} \exists \mu_i \in \Delta(\Theta_{-i}\times M_{-i}) \text{ s.t.} \\ (1)\ \mu_i(\theta_{-i},m_{-i})>0 \Rightarrow m_{-i}\in S_{-i}^k[\mathscr{M}](\theta_{-i}) \\ (2)\ m_i \in \arg\max_{m_i'} \sum_{\theta_{-i},m_{-i}} \mu_i(\theta_{-i},m_{-i}) u_i(g(m_i',m_{-i}),(\theta_i,\theta_{-i})) \end{array} \right\}.$$

We let

$$S_i[\mathscr{M}](\theta_i) = \bigcap_{k\geq 0} S_i^k[\mathscr{M}](\theta_i).$$

We refer to $S_i[\mathscr{M}](\theta_i)$ as the *rationalizable* messages of type θ_i of agent i in mechanism $\mathscr{M}$. This incomplete information version of rationalizability is studied in Battigalli (1999), Battigalli and Siniscalchi (2003), and Bergemann and Morris (2008). A standard and well known duality argument implies that this solution concept is equivalent to iterated deletion of ex post strictly dominated strategies.

The set $S_i[\mathscr{M}](\theta_i)$ is the set of messages that type θ_i might send consistent with knowing that his payoff type is θ_i, common knowledge of rationality, and the set of possible payoff types of the other players, but with no restrictions on his beliefs and higher order beliefs about other types. Equivalently, it is the set of messages that might be played in any equilibrium on any type space by a type of player i with payoff type θ_i and any possible beliefs and higher order beliefs about others' payoff types. In Section 6.4, we report a formal argument confirming this interpretation. In the body of the paper, we work directly with this solution concept.

2.3 *Separability*

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 possible types Θ_j. Thus writing $\Psi_{-i}=\{\Psi_j\}_{j\neq i}$ for a profile of subsets of i's opponents, we define

$$\mathscr{R}_i(\theta_i,\Psi_{-i}) = \{R\in\mathscr{R} \mid R = R_{\theta_i,\lambda_i} \text{ for some } \lambda_i\in\Delta(\Psi_{-i})\}.$$

We adopt the convention that if for some $j\neq i$, $\Psi_j=\varnothing$, then $\mathscr{R}_i(\theta_i,\Psi_{-i})=\varnothing$. Now suppose we observe i's preferences over lotteries and know that i assigns probability 1 to his opponents' type profile θ_{-i} being an element of Ψ_{-i}. What can we deduce about i's type? We say that Ψ_{-i} *separates* Ψ_i if, whatever those realized preferences, we can rule out at least one possible type of i.

Definition 1 (Separation). The type set profile Ψ_{-i} *separates* Ψ_i if

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

We are interested in a process by which we iteratively delete type sets of each agent that are separated by some type set profile of his opponents. Thus writing Ξ_i^k for the kth level inseparable sets of player i, we have

$$\Xi_i^0 = 2^{\Theta_i} \tag{1}$$

and

$$\Xi_i^{k+1} = \{\Psi_i \in \Xi_i^k \mid \Psi_{-i} \text{ does not separate } \Psi_i, \text{ for some } \Psi_{-i} \in \Xi_{-i}^k\}, \tag{2}$$

and a (finite) limit type set profile is defined by

$$\Xi_i^* = \bigcap_{k \geq 0} \Xi_i^k. \tag{3}$$

Finally, we say that two types are pairwise inseparable if they cannot be iteratively separated in this way.

Definition 2 (Pairwise Inseparability). Types θ_i and θ_i' are *pairwise inseparable*, written $\theta_i \sim \theta_i'$, if $\{\theta_i, \theta_i'\} \in \Xi_i^*$.

Note that the relation $\sim$ is reflexive and symmetric by construction, but is not necessarily transitive. The following "fixed point" characterization of pairwise inseparability is useful in the analysis that follows. Let $\Xi = (\Xi_i)_{i=1}^I \in \times_{i=1}^I 2^{\Theta_i}$ be a profile of type sets for each agent.

Definition 3 (Mutual Inseparability). 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.

Lemma 3. *Types θ_i and θ_i' are pairwise inseparable if and only if there exist mutually inseparable $\Xi = (\Xi_i)_{i=1}^I$ and $\Psi_i \in \Xi_i$ with $\{\theta_i, \theta_i'\} \subseteq \Psi_i$.*

Proof. (if) Suppose there exist $\widehat{\Xi} = (\widehat{\Xi}_i)_{i=1}^I$ and $\Psi_i \in \widehat{\Xi}_i$ with $\{\theta_i, \theta_i'\} \subseteq \Psi_i$. We claim that

$$\{\Psi_i \mid \Psi_i \subseteq \Psi_i' \text{ and } \Psi_i' \in \widehat{\Xi}_i \text{ for some } \Psi_i'\} \subseteq \Xi_i^k$$

for each $k = 0, 1, \ldots$. The claim holds for $k = 0$ by definition. Suppose the claim holds for arbitrary k and suppose that $\Psi_i \subseteq \Psi_i'$ and $\Psi_i' \in \widehat{\Xi}_i$. Because $\widehat{\Xi}$ is mutually inseparable, there exists $\Psi_{-i} \in \widehat{\Xi}_{-i} \subseteq \Xi_i^k$ such that Ψ_{-i} does not separate Ψ_i'. By the definition of separation, since $\Psi_i \subseteq \Psi_i'$, Ψ_{-i} does not separate Ψ_i. So $\Psi_i \in \Xi_i^{k+1}$ and

$$\{\theta_i, \theta_i'\} \subseteq \Psi_i \in \Xi_i^* = \bigcap_{k \geq 0} \Xi_i^k.$$

(only if) Observe that $\Xi_i^{k+1} \subseteq \Xi_i^k$ for each $k = 0, 1, \ldots$ by construction. Thus $(\Xi_i^*)_{i=1}^I$ is mutually inseparable. Thus if $\theta_i \sim \theta_i'$, there exists a mutually inseparable Ξ^* with $\{\theta_i, \theta_i'\} \in \Xi_i^*$. □

3. An environment with interdependent values for a single good

We consider a quasi-linear environment with a single good with interdependent values to illustrate the notion of separability. There are I agents and agent i's payoff type is $\theta_i \in [0,1]$. If the type profile is θ, agent i's valuation of an object is given by

$$v_i(\theta_i, \theta_{-i}) = \theta_i + \gamma \sum_{j \neq i} \theta_j,$$

with $\gamma \in \mathbb{R}_+$. The parameter γ measures the amount of interdependence in valuations: the case of private values is given by $\gamma = 0$ and the case of pure common values is $\gamma = 1$. The net utility of agent i depends on his probability y_i of receiving the object and the monetary transfer t_i:

$$u_i(\theta, y_i, t_i) = \left(\theta_i + \gamma \sum_{j \neq i} \theta_j\right) y_i - t_i.$$

We determine the conditions for separability of types in this preference environment.[6]

Type set profile Ψ_{-i} separates Ψ_i if, knowing i's preferences and knowing that he is sure that others' type profile is Ψ_{-i}, we can always rule out some θ_i. In this example, because the utility function u_i is linear in the monetary transfer for all types and all agents, separability must come from different valuations of the object. For a given type set profile Ψ_{-i} of all but i, we can identify the set of possible (expected) valuations of agent i with type θ_i by writing

$$\begin{aligned} V_i(\theta_i, \Psi_{-i}) &= \left\{ v_i \in \mathbb{R}_+ \mid \exists \lambda_i \in \Delta(\Psi_{-i}) \text{ s.t. } v_i = \theta_i + \gamma \sum_{\theta_{-i} \in \Psi_{-i}} \lambda_i(\theta_{-i}) \sum_{j \neq i} \theta_j \right\} \\ &= \left[\theta_i + \gamma \sum_{j \neq i} \min \Psi_j, \theta_i + \gamma \sum_{j \neq i} \max \Psi_j \right]. \end{aligned} \tag{4}$$

Now Ψ_{-i} separates Ψ_i if and only if

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

This is equivalent to requiring that

$$V_i(\max \Psi_i, \Psi_{-i}) \cap V_i(\min \Psi_i, \Psi_{-i}) = \varnothing.$$

By (4), this holds if and only if

$$\max \Psi_i + \gamma \sum_{j \neq i} \min \Psi_j > \min \Psi_i + \gamma \sum_{j \neq i} \max \Psi_j.$$

[6]This example has a continuum of types and a continuum of deterministic monetary allocations while the general model is defined for a finite number of types and pure outcomes. We could rewrite the example and the corresponding results without loss in the finite setting. With a finite model, integer problems would need to be taken into account; in particular, the exact value of the critical threshold for moderate interdependence would depend on the size of the grid. But as the grid becomes finer, the critical thresholds converge to the ones of the continuum example here.

We can rewrite the inequality as

$$\max \Psi_i - \min \Psi_i > \gamma \sum_{j \neq i} (\max \Psi_j - \min \Psi_j).$$

Thus Ψ_{-i} separates Ψ_i if and only if the difference between the smallest and the largest elements in the set Ψ_i is larger than the weighted sum of the differences of the smallest and the largest elements in the remaining sets Ψ_j for all $j \neq i$. Conversely, Ψ_{-i} does not separate Ψ_i if the above inequality is reversed, i.e.,

$$\max \Psi_i - \min \Psi_i \leq \gamma \sum_{j \neq i} (\max \Psi_j - \min \Psi_j). \tag{5}$$

Now we can identify the kth level inseparable sets, described in (1)–(3), for our example. We have

$$\Xi_i^0 = 2^{[0,1]}$$

and, by (5),

$$\Xi_i^k = \left\{ \Psi_i \in \Xi_i^k \mid \max \Psi_i - \min \Psi_i \leq \gamma \sum_{j \neq i} \max_{\Psi_j \in \Xi_j^k} (\max \Psi_j - \min \Psi_j) \right\}.$$

Now by induction, we have

$$\Xi_i^{k+1} = \left\{ \Psi_i \mid \max \Psi_i - \min \Psi_i \leq (\gamma(I-1))^k \right\}.$$

Thus if $\gamma(I-1) < 1$, Ξ_i^* consists of singletons, $\Xi_i^* = (\{\theta_i\})_{\theta_i \in [0,1]}$, while if $\gamma(I-1) \geq 1$, Ξ_i^* consists of all subsets, $\Xi_i^* = 2^{[0,1]}$.

Thus if $\gamma < 1/(I-1)$, so that interdependence is not too large, every distinct pair of types are pairwise separable. If $\gamma \geq 1/(I-1)$, every pair of types are pairwise inseparable. We note that the linear structure of the valuations v_i leads to the strong converse result. But the example illustrates the general principle that pairwise separability corresponds to not too much interdependence. We state a more general result about the relationship between pairwise separability and not too much interdependence in Section 6.2. We also note that the argument surrounding the pairwise separability result relies on the boundedness of the payoff type space. In particular if $\Theta_i = \mathbb{R}$, then pairwise separability can only be achieved in the case of of pure private values, i.e. $\gamma = 0$.

Our later results show that if $\gamma \geq 1/(I-1)$, no social choice function (except for a constant one) is robustly virtually implementable; but if $\gamma < 1/(I-1)$, any ex post incentive compatible allocation can be robustly virtually implemented. One can construct generalized VCG payments such that efficient allocation is ex post incentive compatible in this environment if $\gamma \leq 1$. Thus the efficient allocation is robustly virtually implementable if and only if $\gamma < 1/(I-1)$.

Our result on robust virtual implementation in this environment contrasts with what happens with standard Bayesian implementation. Suppose we assume there is common knowledge of a common prior on the set of payoff types $[0,1]^I$. Suppose first that agents'

types are drawn independently. Then each type has different expected valuations of the object and can easily be separated. Even if priors are not independent, for a "typical" choice of prior, the measurability condition of Abreu and Matsushima (1992b) and Bayesian virtual implementation are possible as long as incentive compatibility conditions are satisfied. Ex post incentive compatibility (and thus Bayesian incentive compatibility for any prior) is satisfied by the efficient allocation if $\gamma \leq 1$.

4. Strategic distinguishability

4.1 *Main result*

Two payoff types are strategically distinguishable if there exists a mechanism where the rationalizable actions of those payoff types are disjoint. Thus they are strategically indistinguishable if they have a rationalizable action in common in every mechanism.

Definition 4 (Strategically Indistinguishable). Types θ_i and θ_i' are *strategically indistinguishable* if $S_i[\mathcal{M}](\theta_i) \cap S_i[\mathcal{M}](\theta_i') \neq \varnothing$ for every $\mathcal{M}$.

The notion of strategic indistinguishability is related to the idea of incentive compatibility in the context of information revelation in a mechanism. The difference between distinguishability and incentive compatibility arises from the two central features of strategic indistinguishability. First, we say that two payoff types can be strategically distinguished if there exists *some* mechanism and hence *some* outcome function for which the types have disjoint rationalizable actions. In contrast, the analysis of incentive compatibility is typically concerned with a specific mechanism and hence a specific outcome function. Second, strategic distinguishability requires that the two payoff types display disjoint rationalizable actions for *all possible* beliefs and higher order beliefs. In contrast, the analysis of incentive compatibility is typically concerned with a fixed and common prior belief of the agents.

The characterization of strategic indistinguishability is the key result in our characterization of robust virtual implementation.

Theorem 1 (Equivalence). *Types θ_i and θ_i' are strategically indistinguishable if and only if they are pairwise inseparable.*

This result is proved in two parts. First, Proposition 1 shows that under any finite mechanism, if θ_i and θ_i' are pairwise inseparable, then the intersection of the set of rationalizable messages for θ_i and θ_i' is always non-empty. This observation follows easily from our definitions.

Proposition 1. *If θ_i and θ_i' are pairwise inseparable ($\theta_i \sim \theta_i'$), then*

$$S_i[\mathcal{M}](\theta_i) \cap S_i[\mathcal{M}](\theta_i') \neq \varnothing$$

in any mechanism $\mathcal{M}$.

PROOF. By Lemma 3, if $\theta_i \sim \theta_i'$, there exists a mutually inseparable Ξ with $\{\theta_i, \theta_i'\} \subseteq \Psi_i^* \in \Xi_i$.

Now fix any mechanism $\mathscr{M}$. We show, by induction on k, that for each k, i, and $\Psi_i \in \Xi_i$, there exists $m_i^k(\Psi_i) \in M_i$ such that $m_i^k(\Psi_i) \in S_i^k[\mathscr{M}](\widetilde{\theta}_i)$ for each $\widetilde{\theta}_i \in \Psi_i$. This is true by definition for $k = 0$. Suppose that it is true for k. Now fix any i and $\Psi_i \in \Xi_i$. Since Ξ is mutually inseparable,

$$\text{there exists } \Psi_{-i} \in \Xi_{-i}, R, \text{ and, for each } \widetilde{\theta}_i \in \Psi_i, \lambda_i^{\widetilde{\theta}_i} \in \Delta(\Psi_{-i}) \text{ such that } R_{\widetilde{\theta}_i, \lambda_i^{\widetilde{\theta}_i}} = R.$$

Now let $m_i^{k+1}(\Psi_i)$ be any optimal message of agent i when he believes that his opponents will send the message profile $m_{-i}^k(\Psi_{-i})$ with probability 1 and has beliefs $\lambda_i^{\widetilde{\theta}_i}$ about the type profile of his opponents, i.e.,

$$m_i^{k+1}(\Psi_i) \in \arg\max_{m_i'} \sum_{\theta_{-i}} \lambda_i^{\widetilde{\theta}_i}(\theta_{-i}) u_i(g(m_i', m_{-i}^k(\Psi_{-i})), (\widetilde{\theta}_i, \theta_{-i})).$$

By construction, $m_i^{k+1}(\Psi_i) \in S_i^{k+1}[\mathscr{M}](\widetilde{\theta}_i)$ for all $\widetilde{\theta}_i \in \Psi_i$.

By the finiteness of the mechanism, there exists K such that $S_i^k[\mathscr{M}](\widetilde{\theta}_i) = S_i[\mathscr{M}](\widetilde{\theta}_i)$ for all i, $\widetilde{\theta}_i$ and $k \geq K$. Thus for each $\Psi_i \in \Xi_i$, there exists $m_i(\Psi_i) \in M_i$ such that $m_i(\Psi_i) \in S_i[\mathscr{M}](\widetilde{\theta}_i)$ for each $\widetilde{\theta}_i \in \Psi_i^*$. Thus there exists $m_i \in S_i[\mathscr{M}](\theta_i) \cap S_i[\mathscr{M}](\theta_i')$. □

The second part of the proof of the theorem is the converse result.

PROPOSITION 2 (Existence of Maximally Revealing Mechanism). *There exists $\mathscr{M}^*$ such that $\theta_i \nsim \theta_i' \Rightarrow S_i[\mathscr{M}^*](\theta_i) \cap S_i[\mathscr{M}^*](\theta_i') = \varnothing$.*

Propositions 1 and 2 immediately imply Theorem 1. Proposition 2 is proved by the explicit construction of a mechanism that leads every pair of distinguishable types to choose different messages. We refer to the specific mechanism as the "maximally revealing mechanism," and spend the rest of this section describing its construction and finding its properties.

4.2 *The maximally revealing mechanism*

We construct a mechanism that works for any environment. In the canonical mechanism, each agent is given K simultaneous opportunities to select a preferred allocation from a given "test set" of allocations. For each opportunity k to select a preferred allocation, with $k = 1, \ldots, K$, the agent is asked to report a profile of possible choices by the remaining agents in the opportunities preceding the kth opportunity. If the report of the agent at opportunity k matches the choices of the other agents in the opportunities below k, then he is given the right to choose a preferred allocation. On the other hand, if his report fails to replicate the choices of the other agents in the opportunities before k, then the designer simply selects the central lottery $\bar{y}$. While the mechanism is entirely static, it requires each agent to make a series of choices, each one contingent on the choices of the other agents. In particular, by asking the agent at opportunity k to match his report with the choices of the other agents at the opportunities before k, we

introduce an inductive structure into the series of choices by each agent. We therefore refer to the kth opportunity as the kth stage or kth step of the mechanism even though the mechanism itself is entirely static.

The central aspect of the inductive structure of the choice mechanism is that it allows us to analyze the behavior of the agent in the mechanism in terms of the iterative elimination of dominated strategies. The precise construction of the choice mechanism is based on two central concepts, the notion of a test set and the notion of an augmentation of a given mechanism. A test set gives each agent a finite set of choices and the choice behavior by the agent allows us to distinguish between different types of the agent. The construction of the set of test allocations relies on a few critical implications of our notion of separation. In turn, the notion of an augmentation permits us to show that we can always construct a more informative mechanism on the basis of a given mechanism.

4.2.1 *A class of maximally revealing mechanisms* Fix a finite "test set" of lotteries Y^*. A maximally revealing mechanism offers each agent i a series of K opportunities to select a preferred allocation from Y^*. The set of messages for each agent in a maximally revealing mechanism is defined as follows. Let $M_i^0 = \{\bar{m}_i^0\}$ and inductively define

$$M_i^{k+1} = M_i^k \times M_{-i}^k \times Y^*.$$

Thus $M_i^0 = \{\bar{m}_i^0\}$, $M_i^1 = \{\bar{m}_i^0\} \times M_{-i}^0 \times Y^*$, $M_i^2 = \{\bar{m}_i^0\} \times M_{-i}^0 \times Y^* \times M_{-i}^1 \times Y^*$, and so on. The message m_i^{k+1} of agent i in stage $k+1$ thus reiterates his message from step k and announces a possible message profile of the remaining agents in step k. Due to the inductive structure of the messages, we can write a typical element $m_i^k \in M_i^k$ as a list of the form

$$m_i^k = \{m_i^0, r_i^1, y_i^1, r_i^2, y_i^2, \ldots, r_i^k, y_i^k\}, \tag{6}$$

with $m_i^0 = \bar{m}_i^0$ and each $r_i^k \in M_{-i}^{k-1}$ and each $y_i^k \in Y^*$. The entry r_i^k constitutes the report of agent i regarding the message of the other agents in the previous stage $k-1$. The message set of agent i is then given by M_i^K.

The outcome function in the revealing mechanism is defined as

$$g^{K,\varepsilon}(m) \triangleq \bar{y} + \left(\frac{1-\varepsilon^K}{1-\varepsilon}\right)\frac{1}{I}\left(\sum_{k=1}^{K}\varepsilon^{k-1}\sum_{i=1}^{I}\mathbb{I}(r_i^k, m_{-i}^{k-1})(y_i^k - \bar{y})\right) \tag{7}$$

for some $\varepsilon > 0$, where $\mathbb{I}$ is the indicator function:

$$\mathbb{I}(r_i^k, m_{-i}^{k-1}) \triangleq \begin{cases} 1 & \text{if } r_i^k = m_{-i}^{k-1} \\ 0 & \text{otherwise.} \end{cases}$$

For a given $\varepsilon > 0$ and positive integer K, we refer to the associated revealing mechanism as

$$\mathscr{M}_\varepsilon^K \triangleq (M^K, g_\varepsilon^K).$$

In words, the mechanism has K stages. In each stage k, an agent is asked to announce a stage $k-1$ profile of messages he thinks his opponents might have sent and, with

positive probability, gets to pick a lottery from Y^*. Lotteries from early stages are much more likely to be chosen than lotteries from later stages. We can now analyze how the series of messages can iteratively and interactively identify the types of each agent.

4.2.2 *Characterizing rationalizable behavior for small ε* For sufficiently small $\varepsilon > 0$, an agent's choice of a message at the kth stage is independent of what messages he thinks others will send at stage k and higher and thus also independent of K, the total number of stages of messages that will be sent. We first propose an inductive characterization of the set of types of player i who could possibly send kth stage message m_i^k, and denote this set by $\bar{\Theta}_i^k(m_i^k)$. We then verify with Lemmas 6 and 8 that our proposed inductive characterization of rationalizable messages is correct for sufficiently small ε.

Write $B_i^{Y^*}(\theta_i, \lambda_i)$ for agent i's most preferred lotteries in the set Y^* if he has payoff type θ_i and beliefs $\lambda_i \in \Delta(\Theta_{-i})$ and (with a minor abuse of notation) let $B_i^{Y^*}(\theta_i, \Psi_{-i})$ be agent i's possible most preferred lotteries if he has payoff type θ_i and assigns probability 1 to his opponents having types in Ψ_{-i}, so that

$$B_i^{Y^*}(\theta_i, \lambda_i) \triangleq \{y \in Y^* \mid y \, R_{\theta_i, \lambda_i} \, y' \text{ for all } y' \in Y^*\}$$

and

$$B_i^{Y^*}(\theta_i, \Psi_{-i}) \triangleq \bigcup_{\lambda_i \in \Delta(\Psi_{-i})} B_i^{Y^*}(\theta_i, \lambda_i).$$

We adopt the convention that if $\Psi_j = \varnothing$ for some $j \neq i$, then $B_i^{Y^*}(\theta_i, \varnothing) = \varnothing$ as well.

Let $\bar{\Theta}_i^1(m_i^1)$ be the set of types of player i who could possibly send first stage message m_i^1. Since we ignore later stages, this is independent of ε and K. Taking these sets as given, we then find the set $\bar{\Theta}_i^2(m_i^2)$ of types of player i who could possibly send second stage message m_i^2, and so on. We end up with an inductive characterization of the set $\bar{\Theta}_i^k(m_i^k)$ of types of player i who could possibly send kth stage message m_i^k. Thus

$$\bar{\Theta}_i^0(\bar{m}_i^0) \triangleq \Theta_i,$$

and inductively define $\bar{\Theta}_i^{k+1}(m_i^{k+1})$, where we recall that by the inductive description of the message m_i^{k+1} in (6), we have $m_i^{k+1} = (m_i^k, r_i^{k+1}, y_i^{k+1})$:

$$\bar{\Theta}_i^{k+1}(m_i^{k+1}) \triangleq \left\{ \theta_i \in \Theta_i \;\middle|\; \begin{array}{l} \text{(i) } \theta_i \in \bar{\Theta}_i^k(m_i^k) \\ \text{(ii) } \bar{\Theta}_{-i}^k(r_i^{k+1}) \neq \varnothing \\ \text{(iii) } y_i^{k+1} \in B_i^{Y^*}(\theta_i, \bar{\Theta}_{-i}^k(r_i^{k+1})) \end{array} \right\}. \tag{8}$$

The set $\bar{\Theta}_i^k(m_i^k)$ is meant to approximate the set of types of agent i for whom a specific message m_i^k is rationalizable in stage k. In some sense, the set $\bar{\Theta}_i^k(m_i^k)$ is the dual to $S_i^k[\mathscr{M}](\theta_i)$, which describes the set of messages m_i that are rationalizable for a specific type θ_i in stage k. The role of the set $\bar{\Theta}_i^k(m_i^k)$ is to track the information that can be inferred from the choices of messages m_i about the type θ_i of agent i.

The analysis of the limit behavior of $\bar{\Theta}_i^{k+1}(m_i^{k+1})$ is heuristic in the sense that the inductive process assumes the properties (ii) and (iii) in (8). In particular, it is simply

assumed that agent i in stage $k+1$ announces a past message profile of the remaining agents which could have been sent by some type profile of the other agents, and that agent i selects an allocation that is a best response to some belief in stage $k+1$.

We use two preliminary results to establish formally that these sets characterize limit behavior for small ε and large K. The routine proofs are reported in the Appendix. First, we note that for any fixed finite mechanism $\mathcal{M}$, when we iteratively delete messages that are not best responses, they are uniformly worse responses, i.e., there exists $\eta_{\mathcal{M}} > 0$ such that each of those deleted messages is not even an $\eta_{\mathcal{M}}$-best response.

LEMMA 4 (Uniformly Worse Responses). *For any mechanism $\mathcal{M}$, there exists $\eta_{\mathcal{M}} > 0$ such that if $m_i \in S_i^k[\mathcal{M}](\theta_i)$, $m_i \notin S_i^{k+1}[\mathcal{M}](\theta_i)$, and $\mu_i \in \Delta(\Theta_{-i} \times M_{-i})$ satisfies*

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

then there exists $\bar{m}_i$ such that

$$\sum_{\theta_{-i}, m_{-i}} \mu_i(\theta_{-i}, m_{-i}) u_i(g^*(\bar{m}_i, m_{-i}), (\theta_i, \theta_{-i})) > \sum_{\theta_{-i}, m_{-i}} \mu_i(\theta_{-i}, m_{-i}) u_i(g^*(m_i, m_{-i}), (\theta_i, \theta_{-i})) + \eta_{\mathcal{M}}.$$

Second, we use the uniform lower bound in stating a key result about "augmenting" mechanisms. We use this "augmentation lemma" in the construction of both the maximally revealing mechanism (in this section) and the canonical mechanism for robust virtual implementation (in the next section). For each player i, fix finite message sets M_i^0 and M_i^1 and let $M_i = M_i^0 \times M_i^1$. Fix $g^0 : M^0 \to Y$, $g^1 : M^1 \to Y$, and $g^+ : M \to Y$. Fix $\pi^0, \pi^1, \pi^+ \geq 0$ with $\pi^0 + \pi^1 + \pi^+ = 1$ and let $g : M \to Y$ be defined by

$$g(m) \triangleq \pi^0 g^0(m^0) + \pi^1 g^1(m^1) + \pi^+ g^+(m).$$

We now consider the mechanism

$$\mathcal{M}^0 \triangleq ((M_i^0)_{i=1}^I, g^0)$$

and the augmented mechanism

$$\mathcal{M} \triangleq ((M_i)_{i=1}^I, g).$$

We recall that the constant $C > 0$ is a finite upper bound on the difference in payoffs across all agents and all pairs of lotteries y and y', which we established earlier in Lemma 2.

LEMMA 5 (Augmentation). *If $\pi^+ C \leq \pi^0 \eta_{\mathcal{M}^0}$, then*

$$(m_i^0, m_i^1) \in S_i[\mathcal{M}](\theta_i) \Rightarrow m_i^0 \in S_i[\mathcal{M}^0](\theta_i).$$

The lemma states that if the weight π^0 put on the original payoff function g^0 in the augmented mechanism is much larger than the weight π^+ put on the other component of the mechanism at which m^0 affects the allocation, then any rationalizable message in the augmented mechanism must entail sending a message m_i^0 that was rationalizable in the original mechanism.

We now show that these choices are indeed the result of iterative elimination of strictly dominated strategies. More precisely, we verify that $\bar{\Theta}_i^k(m_i^k)$ is an upper bound on the set of types who could send kth stage message m_i^k in any $\mathscr{M}_\varepsilon^k$ for sufficiently small ε.

LEMMA 6 (Limit). *Suppose that $B_i^{Y^*}(\theta_i,\lambda_i)\neq Y^*$ for each i, θ_i and $\lambda_i\in\Delta(\Theta_{-i})$. Then, for each k, there exists $\bar{\varepsilon}>0$ such that*

$$\{\theta_i\in\Theta_i \mid m_i^k\in S[\mathscr{M}_\varepsilon^k](\theta_i)\}\subseteq\bar{\Theta}_i^k(m_i^k)$$

for all $\varepsilon\leq\bar{\varepsilon}$ and $m_i^k\in M_i^k$.

PROOF. By induction. The claim of holds for $k=0$, since

$$\{\theta_i\in\Theta_i \mid m_i^0\in S[\mathscr{M}_\varepsilon^0](\theta_i)\}=\Theta_i=\bar{\Theta}_i^0(m_i^0).$$

Now suppose that the claim holds for k. Thus there exists $\bar{\varepsilon}_k>0$ such that

$$\{\theta_i\in\Theta_i \mid m_i^k\in S[\mathscr{M}_\varepsilon^k](\theta_i)\}\subseteq\bar{\Theta}_i^k(m_i^k) \text{ for all } \varepsilon\leq\bar{\varepsilon}_k \text{and } m_i^k\in M_i^k.$$

Now observe that $\mathscr{M}_\varepsilon^{k+1}$ is an augmentation of $\mathscr{M}_\varepsilon^k$ and thus, by Lemma 5, there exists $\bar{\varepsilon}_{k+1}\in(0,\bar{\varepsilon}_k]$ such that for all $\varepsilon\leq\bar{\varepsilon}_{k+1}$,

$$m_i^{k+1}=(m_i^k,r_i^{k+1},y_i^{k+1})\in S[\mathscr{M}_\varepsilon^{k+1}](\theta_i)\Rightarrow m_i^k\in S[\mathscr{M}_\varepsilon^k](\theta_i). \tag{9}$$

Now by the inductive hypothesis, we also have

$$\theta_i\in\bar{\Theta}_i^k(m_i^k). \tag{10}$$

We further observe that $m_i^{k+1}\in S[\mathscr{M}_\varepsilon^{k+1}](\theta_i)$ also implies there must exist $\mu_i\in\Delta(\Theta_{-i}\times M_{-i}^{k+1})$ such that (1):

$$\mu_i(\theta_{-i},m_{-i}^{k+1})>0\Rightarrow\ m_j^{k+1}\in S[\mathscr{M}_\varepsilon^{k+1}](\theta_j) \text{ for each } j\neq i$$

and (2):

$$m_i^{k+1}\in\operatorname*{arg\,max}_{\bar{m}_i^{k+1}\in M_i^{k+1}}\sum_{\theta_{-i},m_{-i}^{k+1}}\mu_i(\theta_{-i},m_{-i}^{k+1})[u_i(g^{k+1,\varepsilon}(\bar{m}_i^{k+1},m_{-i}^{k+1}),(\theta_i,\theta_{-i}))].$$

But note that (r_i^{k+1},y_i^{k+1}), the last components of m_i^{k+1}, affects only one additively separable component of the above expression. In particular, (r_i^{k+1},y_i^{k+1}) must maximize

$$\sum_{\theta_{-i},m_{-i}^{k+1}}\mu_i(\theta_{-i},m_{-i}^{k+1})\mathbb{I}(r_i^{k+1},m_{-i}^k)\big(u_i(y_i^{k+1},(\theta_i,\theta_{-i}))-u_i(\bar{y},(\theta_i,\theta_{-i}))\big), \tag{11}$$

which we can rewrite as

$$\sum_{\theta_{-i}} \sum_{\{m_{-i}^{k+1} | m_{-i}^{k} = r_i^{k+1}\}} \mu_i(\theta_{-i}, m_{-i}^{k+1}) \big(u_i(y_i^{k+1}, (\theta_i, \theta_{-i})) - u_i(\bar{y}, (\theta_i, \theta_{-i}))\big).$$

In particular, the later expression is zero if

$$\mu_i(r_i^{k+1}) \triangleq \sum_{\theta_{-i}} \sum_{\{m_{-i}^{k+1} | m_{-i}^{k} = r_i^{k+1}\}} \mu_i(\theta_{-i}, m_{-i}^{k+1}) = 0.$$

But if $\mu_i(r_i^{k+1}) > 0$ and $y_i^{k+1} \in B_i^{Y^*}(\theta_i, \lambda_i)$, where

$$\lambda_i(\theta_{-i}) = \frac{\sum_{\{m_{-i}^{k+1} | m_{-i}^{k} = r_i^{k+1}\}} \mu_i(\theta_{-i}, m_{-i}^{k+1})}{\sum_{\theta'_{-i}} \sum_{\{m_{-i}^{k+1} | m_{-i}^{k} = r_i^{k+1}\}} \mu_i(\theta'_{-i}, m_{-i}^{k+1})},$$

then (11) must be strictly positive, by the premise of the lemma. Thus we must have (r_i^{k+1}, y_i^{k+1}) chosen such that $\mu_i(r_i^{k+1}) > 0$ and $y_i^{k+1} \in B_i^{Y^*}(\theta_i, \lambda_i)$. Now $\mu_i(r_i^{k+1}) > 0$, (9), and the inductive hypothesis imply that

$$\bar{\Theta}_{-i}^{k}(r_i^{k+1}) \neq \varnothing \tag{12}$$

and

$$\lambda_i \in \Delta(\bar{\Theta}_{-i}^{k}(r_i^{k+1})) \quad \text{and} \quad y_i^{k+1} \in B_i^{Y^*}(\theta_i, \lambda_i). \tag{13}$$

To wit, by the construction of the revealing mechanism (see (7)), the lottery y_i^{k+1} specified in (13) only affects the (expected) payoff of agent i when $r_i^{k+1} = m_{-i}^{k}$. It follows that y_i^{k+1} should be a best reply to some belief conditioned on the event that $r_i^{k+1} = m_{-i}^{k}$.

Now (10), (12), and (13) together imply that any message $m_i^{k+1} \in S[\mathscr{M}_\varepsilon^{k+1}](\theta_i)$ satisfies the three requirements in the construction of $\bar{\Theta}_i^{k+1}(r_i^{k+1})$ in (8) and hence for any $m_i^{k+1} \in S[\mathscr{M}_\varepsilon^{k+1}](\theta_i)$ we have $\theta_i \in \bar{\Theta}_i^{k+1}(m_i^{k+1})$. □

4.3 *Constructing a rich enough test set*

Finally, we show that we can choose the "test set" Y^* to be sufficiently large so that Lemma 6 implies that, for sufficiently small $\varepsilon > 0$ and sufficiently large K, the members of any pair of mutually separable types are sending distinct messages in the (K, ε) revealing mechanism.

Lemma 7 (Existence of Finite Test Set). *There exists a finite test set $Y^* \subseteq Y$ such that*

(i) for each i, θ_i, and $\lambda_i \in \Delta(\Theta_{-i})$, $B_i^{Y^*}(\theta_i, \lambda_i) \neq Y^*$

(ii) for each i, Ψ_i, and Ψ_{-i}, if Ψ_{-i} separates Ψ_i, then for each $\theta_i \in \Psi_i$ and $\lambda_i \in \Delta(\Psi_{-i})$, there exists $\theta'_i \in \Psi_i$ such that

$$B_i^{Y^*}(\theta_i, \lambda_i) \cap B_i^{Y^*}(\theta'_i, \Psi_{-i}) = \varnothing.$$

The proof of Lemma 7 is in the Appendix. The proof of Proposition 2 is completed by the following lemma, establishing that the sets $\bar{\Theta}_i^k$ are closely related to kth level inseparable sets Ξ_i^k, as defined in (1)–(3).

LEMMA 8. *For all i, all k, and all $m_i^k \in M_i^k$, $\bar{\Theta}_i^k(m_i^k) \subseteq \Psi_i$ for some $\Psi_i \in \Xi_i^k$.*

PROOF. By induction. The claim holds for $k=0$ by definition. Suppose for all $m_{-i}^k \in M_{-i}^k$ we have $\bar{\Theta}_{-i}^k(m_{-i}^k) \subseteq \Psi_i$ for some $\Psi_i \in \Xi_i^k$. Now fix any $m_i^{k+1} = (m_i^k, r_i^{k+1}, y_i^{k+1}) \in M_i^{k+1}$. If $\bar{\Theta}_i^{k+1}(m_i^{k+1}) = \varnothing$, then we are done as the empty set is included in every $\Psi_i \neq \varnothing$. If $\bar{\Theta}_i^{k+1}(m_i^{k+1}) \neq \varnothing$, then we let $\Psi_i = \bar{\Theta}_i^{k+1}(m_i^{k+1})$ and $\Psi_{-i} = \bar{\Theta}_{-i}^k(r_i^{k+1})$. Lemma 7(i) ensures that for every θ_i and λ_i, there exist $y, y' \in \widetilde{Y}$ such that $y \; P_{\theta_i,\lambda_i} \; y'$. Thus any best response involves setting r_i^{k+1} equal to some m_{-i}^k to which he assigns positive probability and choosing a strictly preferred lottery. By our inductive assumption, $\Psi_{-i} \in \Xi_{-i}^k$. Now suppose Ψ_{-i} separates Ψ_i and fix $\theta_i \in \Psi_i$. By Lemma 7(ii), there exists $\theta_i' \in \Psi_i$ such that $y_i^{k+1} \notin B_i^{Y^*}(\theta_i', \Psi_{-i})$. Thus $\theta_i' \notin \bar{\Theta}_i^{k+1}(m_i^{k+1})$, a contradiction. We conclude that Ψ_{-i} does not separate Ψ_i. □

5. ROBUST VIRTUAL IMPLEMENTATION

In this section, we use the notions of strategic distinguishability and the maximally revealing mechanism to establish necessary and sufficient conditions for robust virtual implementation. Virtual implementation of a social choice function requires a mechanism such that the desired outcomes are realized with probability arbitrarily close to 1 (see Abreu and Matsushima 1992b,c). Robust implementation requires implementation of a social choice function depending on agents' "payoff types" independent of their beliefs and higher order beliefs about others' payoff types (see Bergemann and Morris 2008, forthcoming). Our definition of robust virtual implementation is the natural one incorporating both these notions.

5.1 *Definitions*

Write $\|y - y'\|$ for the rectilinear norm between a pair of lotteries y and y', i.e.,

$$\|y - y'\| \triangleq \sum_{x \in X} |y(x) - y'(x)|.$$

DEFINITION 5 (Robust ε-Implementation). The mechanism $\mathcal{M}$ *robustly ε-implements* the social choice function f if

$$m \in S[\mathcal{M}](\theta) \Rightarrow \|g(m) - f(\theta)\| \leq \varepsilon.$$

The social choice function f is robustly ε-implementable if there exists a mechanism $\mathcal{M}$ that robustly ε-implements f.

We can now define the notion of robust virtual implementation.

DEFINITION 6 (Robust Virtual Implementation). The social choice function f is *robustly virtually implementable* if, for every $\varepsilon > 0$, f is robustly ε-implementable.

The relevant incentive compatibility condition required for our robust problem is ex post incentive compatibility.

DEFINITION 7 (EPIC). The social choice function f satisfies *ex post incentive compatibility* (EPIC) if, for all i, θ_i, θ_{-i}, and θ_i',

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

Now, "robust measurability" requires that if θ_i is pairwise inseparable from θ_i', then the social choice function must treat the two types the same. This condition is the robust analogue of the measurability condition in Abreu and Matsushima (1992c) as we formally establish in Section 6.1.

DEFINITION 8 (Robust Measurability). The social choice function f is *robustly measurable* if $\theta_i \sim \theta_i' \Rightarrow f(\theta_i,\theta_{-i}) = f(\theta_i',\theta_{-i})$ for all θ_{-i}.

5.2 *Necessity*

It is well known from the literature on virtual Bayesian implementation (e.g., Abreu and Matsushima 1992c) that the relaxation to virtual implementation does not relax incentive compatibility conditions by a standard compactness argument.[7]

THEOREM 2 (Necessity). *If f is robustly virtually implementable, then f is ex post incentive compatible and robustly measurable.*

PROOF. We first establish ex post incentive compatibility. Fix any mechanism $\mathscr{M}$ that robustly ε-implements f. Fix θ_{-i} and $m_{-i} \in S_{-i}^{\mathscr{M}}(\theta_{-i})$. For any $m_i' \in S_i[\mathscr{M}](\theta_i')$, virtual implementation requires

$$\|g(m_i',m_{-i}) - f(\theta_i',\theta_{-i})\| \leq \varepsilon. \tag{14}$$

Now suppose that player i is type θ_i and is convinced that his opponent is type θ_{-i} sending message m_{-i}. Let m_i be any message that is a best response to that belief. Then $m_i \in S_i[\mathscr{M}](\theta_i)$, implying that

$$\|g(m_i,m_{-i}) - f(\theta_i,\theta_{-i})\| \leq \varepsilon. \tag{15}$$

In particular, by the best response property of m_i,

$$u_i(g(m_i,m_{-i}),(\theta_i,\theta_{-i})) \geq u_i(g(m_i',m_{-i}),(\theta_i,\theta_{-i})). \tag{16}$$

Now (14) and Lemma 2 imply

$$\left|u_i(g(m_i',m_{-i}),(\theta_i,\theta_{-i})) - u_i(f(\theta_i',\theta_{-i}),(\theta_i,\theta_{-i}))\right| \leq \tfrac{1}{2}\varepsilon C, \tag{17}$$

and (15) and Lemma 2 imply

$$\left|u_i(g(m_i,m_{-i}),(\theta_i,\theta_{-i})) - u_i(f(\theta_i,\theta_{-i}),(\theta_i,\theta_{-i}))\right| \leq \tfrac{1}{2}\varepsilon C. \tag{18}$$

[7] Dasgupta et al. (1979) and Ledyard (1979) argue in a private value environment that dominant strategy incentive compatibility is implied by Bayesian incentive compatibility for all priors on a fixed type space. In the case of a social choice function, this argument, generalized to interdependent values, shows the necessity of ex post incentive compatibility (see Bergemann and Morris 2005b).

Now combining (16), (17), and (18), we obtain

$$u_i(f(\theta_i,\theta_{-i}),(\theta_i,\theta_{-i}))\geq u_i(f(\theta_i',\theta_{-i}),(\theta_i,\theta_{-i}))-\varepsilon C.$$

But virtual implementation implies that this holds for all $\varepsilon>0$, so we have

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

and this establishes EPIC as a necessary condition.

Next we establish robust measurability. Suppose that f is robustly virtually implementable. Fix any $\varepsilon>0$. Since f is robustly virtually implementable, there exists a mechanism $\mathscr{M}_\varepsilon$ such that

$$m\in S[\mathscr{M}_\varepsilon](\theta)\Rightarrow\|g(m)-f(\theta)\|\leq\varepsilon.$$

Now fix any θ_{-i} and $m_{-i}^\varepsilon\in S_{-i}[\mathscr{M}_\varepsilon](\theta_{-i})$. Also fix any $\theta_i\sim\theta_i'$, so by Proposition 1 there exists

$$m_i^\varepsilon\in S_i[\mathscr{M}_\varepsilon](\theta_i)\cap S_i[\mathscr{M}_\varepsilon](\theta_i').$$

Now $\|g(m_i^\varepsilon,m_{-i}^\varepsilon)-f(\theta_i,\theta_{-i})\|\leq\varepsilon$ and $\|g(m_i^\varepsilon,m_{-i}^\varepsilon)-f(\theta_i',\theta_{-i})\|\leq\varepsilon$. Thus $\|f(\theta_i,\theta_{-i})-f(\theta_i',\theta_{-i})\|\leq 2\varepsilon$. This is true for each $\varepsilon>0$, so $f(\theta_i,\theta_{-i})=f(\theta_i',\theta_{-i})$. □

While we maintain the assumption that the mechanism is finite, the same argument implies the necessity of EPIC and robust measurability if we allow "regular mechanisms" (Abreu and Matsushima 1992c), i.e., mechanisms where best replies always exist for any conjecture over opponents' behavior.

5.3 *Sufficiency*

We first describe the construction of a *canonical mechanism* that we use to establish sufficiency. Our construction follows the logic of Abreu and Matsushima (1992c), which in turn builds on Abreu and Matsushima (1992b). In the mechanism we construct, each agent simultaneously announces (i) a message in the maximally revealing mechanism described above and (ii) L announcements of his payoff type. With probability close to $1/L$, the outcome is chosen according to the agents' lth announcement of their payoff types in part (ii) of their messages. But with small probability, the outcome is chosen according to the maximally revealing mechanism and their part (i) messages. The mechanism then checks to see which agent was the "first" to "lie," in the sense that his lth report of his type is not consistent with the message he sent in the maximally revealing mechanism and no other agent sent an inconsistent message in an "earlier" report. If an agent is not one of the first to lie, then the agent is rewarded. For this part of the mechanism, we need an economic property.

Definition 9 (Economic Property). The *uniform economic property* is satisfied if there exists a profile of lotteries $(z_i)_{i=1}^I$ such that, for each i and θ, $u_i(z_i,\theta)>u_i(\bar{y},\theta)$ and $u_j(\bar{y},\theta)\geq u_j(z_i,\theta)$ for all $j\neq i$.

Under the uniform economic property, there exists a constant c_0 such that

$$u_i(z_i,\theta) > u_i(\bar{y},\theta) + c_0$$

for all i and θ.

In the canonical mechanism, part (i) announcements for the maximally revealing mechanism are made as if the maximally revealing mechanism were being played as a stand-alone mechanism (since the probability of rewards can be chosen sufficiently small). An agent never allows himself to be one of the first to lie: sending a message that ensures that he is not the first to lie (given his beliefs about the others' strategies) always strictly improves his expected payoff, since if the others are telling the truth, truth-telling is a weak best response by ex post incentive compatibility, and if they are lying, for sufficiently large L the reward outweighs the cost of not lying in one stage of the mechanism.

We write $\mathcal{M}^* = ((M_i^*)_{i=1}^I, g^*)$ for the maximally revealing mechanism. We use three numbers in defining the canonical mechanism. The number c_0 is the uniform lower bound on an agent's utility gain from having his uniformly preferred lottery rather than the central lottery. Recall from Lemma 2 that C is an upper bound on payoff differences in the environment, and recall from Lemma 4 that whenever a message is deleted in the iterated deletion process for the maximally revealing mechanism $\mathcal{M}^*$, it is not even an $\eta_{\mathcal{M}^*}$-best response to any conjecture. We use these three numbers c_0, C, and $\eta_{\mathcal{M}^*}$, together with the number of players I, to define two further numbers δ and L that we use in the construction of the canonical mechanism. Choose $\delta > 0$ such that

$$\delta < \frac{\eta_{\mathcal{M}^*}}{C} \tag{19}$$

and an integer L such that

$$L > \frac{IC}{\delta^2 c_0}. \tag{20}$$

Now the message space of the canonical mechanism is

$$M_i = M_i^* \times \overbrace{\Theta_i \times \cdots \times \Theta_i}^{L \text{ times}} = M_i^* \times \Theta_i^L.$$

Thus a typical message is written as $m_i = (m_i^0, m_i^1, \ldots, m_i^L)$, with $m_i^0 \in M_i^*$; $m_i^l \in \Theta_i$ for each $l = 1,\ldots,L$. The idea is that an agent is "supposed" to truthfully report his payoff type in each stage $l = 1,\ldots,L$ and receives a small punishment if he is one of the "first" to report a type that is not consistent with his 0-th message. The small individual rewards and punishments are provided by

$$r_i(m) = \begin{cases} \bar{y} & \text{if } \exists k \in \{1,\ldots,L\} \text{ s.t. } m_i^0 \notin S_i[\mathcal{M}^*](m_i^k) \\ & \qquad \text{and } m_j^0 \in S_j[\mathcal{M}^*](m_j^l)\ \forall j = 1,\ldots,I \text{ and } l = 1,\ldots,k-1 \\ z_i & \text{if otherwise.} \end{cases}$$

(With a slight abuse of notation, we use $r_i(m)$ here to denote rewards, rather than r_i^k as in Section 4.2.1.) Now the outcome function of the canonical mechanism is

$$g(m) = (1-\delta-\delta^2)\frac{1}{L}\sum_{l=1}^{L} f(m^l) + \delta\, g^*(m^0) + \frac{\delta^2}{I}\sum_{i=1}^{I} r_i(m).$$

The mechanism $g(m)$ has three components. The first component, which carries the largest probability, is the social choice function f itself. The appropriate allocation $f(m^l)$ is selected by L replicas, each one of which is chosen with the small probability $1/L$. The second component is the maximally revealing mechanism outcome function g^*, which receives a smaller weight of δ. The third and final component, $r_i(m)$, represents a small reward or punishment. It is designed to give each agent an incentive to replicate in stage l the report issued in the previous stage. It provides a small "punishment" ($\bar{y}$) if player i is the first to report in the message component, m_i^l, a type inconsistent with previous reports; otherwise $r_i(m)$ provides the small "reward" z_i.

THEOREM 3. *Under the uniform economic property, if f satisfies EPIC and robust measurability, then the canonical mechanism $\delta(1+\delta)$ robustly implements f.*

This immediately implies the sufficiency part of our characterization of robust virtual implementation, since we can choose δ arbitrarily close to 0 in the canonical mechanism.

COROLLARY 1 (Sufficiency). *Under the uniform economic property, if f satisfies EPIC and robust measurability, then f is robustly virtually implementable.*

PROOF. To prove the theorem, it is enough to establish that, for each i, $m_i = (m_i^0, m_i^1, \ldots, m_i^L) \in S_i[\mathscr{M}](\theta_i)$ implies that (1) $m_i^0 \in S_i[\mathscr{M}^*](\theta_i)$ and (2) $m_i^0 \in S_i[\mathscr{M}^*](m_i^l)$ for each $l = 1, \ldots, L$. To see why, observe that $m_i^0 \in S_i[\mathscr{M}^*](\theta_i) \cap S_i[\mathscr{M}^*](m_i^l)$ implies θ_i is strategically indistinguishable from m_i^l, which implies, by robust measurability, that $f(m_i^l, m_{-i}^l) = f(\theta_i, m_{-i}^l)$. Since this holds for each i, we have $f(m^l) = f(\theta)$. Since this is true for each l, the mechanism selects $f(\theta)$ with probability at least $1 - \delta - \delta^2$.

Claim (1) above, that $(m_i^0, m_i^1, \ldots, m_i^L) \in S_i[\mathscr{M}](\theta_i) \Rightarrow m_i^0 \in S_i[\mathscr{M}^*](\theta_i)$, follows from Lemma 5 and inequality (19), since m^0 influences the outcome only through weight δ on $g^*(m^0)$ and weight δ^2 on $(1/I)\sum_{i=1}^I r_i(m)$.

We now establish claim (2) above, that $(m_i^0, m_i^1, \ldots, m_i^L) \in S_i[\mathscr{M}](\theta_i) \Rightarrow m_i^0 \in S_i[\mathscr{M}^*](m_i^l)$ for all i and $l = 1, \ldots, L$.

Suppose this claim were false. Then there would exist a smallest l for which the claim fails. Thus there would exist $l^* \in \{1, \ldots, L\}$ such that, for all j, $m_j \in S_j[\mathscr{M}^*](\theta_j) \Rightarrow m_j^0 \in S_j[\mathscr{M}^*](m_j^l)$ for all $1 \leq l < l^*$; but there exist i and $m_i = (m_i^0, m_i^1, \ldots, m_i^L) \in S_i[\mathscr{M}^*](\theta_i)$ with $m_i^0 \notin S_i[\mathscr{M}^*](m_i^{l^*})$. Now fix any conjecture $\mu_i \in \Delta(\Theta_{-i} \times M_{-i})$ with $\mu_i(\theta_{-i}, m_{-i}) > 0 \Rightarrow m_j \in S_j[\mathscr{M}^*](\theta_j)$ for all $j \neq i$. Consider two cases. First, suppose that

$$\mu_i(\theta_{-i}, m_{-i}) > 0 \Rightarrow m_j^0 \in S_j[\mathscr{M}^*](m_j^l) \text{ for all } j \neq i \text{ and } l = 1, \ldots, L. \tag{21}$$

In this case, sending the message

$$\bar{m}_i = (m_i^0, \overbrace{\theta_i, \theta_i, \ldots, \theta_i}^{L \text{ times}})$$

instead of m_i strictly increases i's utility: since he is certain that each agent is reporting a type that is strategically indistinguishable in each of the L stages, EPIC and robust

measurability ensure that his utility does not decrease from truthtelling in the L stages; his utility is unchanged in the maximally revealing mechanism; and his utility is strictly increased in the punishment component. Secondly, i's conjecture μ_i is such that (21) fails. In this case, we can define

$$\widehat{l} = \min\{l \in \{1,\ldots,L\} \mid \exists(\theta_{-i}, m_{-i}) \text{ with } \mu_i(\theta_{-i}, m_{-i}) > 0 \text{ and } m_j^0 \notin S_j[\mathscr{M}^*](m_j^l) \text{ for some } j \neq i\}.$$

Note that $\widehat{l} \geq l^*$. Now sending the message

$$\bar{m}_i = (m_i^0, \overbrace{\theta_i, \theta_i, \ldots, \theta_i}^{\widehat{l} \text{ times}}, m_i^{\widehat{l}+1}, \ldots, m_i^L)$$

instead of m_i strictly increases i's utility by the following argument. Since he is certain that each agent is reporting a type that is strategically indistinguishable in each of the first $\widehat{l}-1$ stages, EPIC and robust measurability ensure that his utility does not decrease from truthtelling in the first $\widehat{l}-1$ stages, and his utility is unchanged in the maximally revealing mechanism. If it turns out that $m_j^0 \in S_j[\mathscr{M}^*](m_j^{\widehat{l}})$ for some $j \neq i$, then i's utility is also not reduced in the $\widehat{l}$-th stage or in the punishment component, but if it turns out that $m_j^0 \notin S_j[\mathscr{M}^*](m_j^{\widehat{l}})$ for all $j \neq i$, then i's utility is reduced in the $\widehat{l}$-th stage by at most $(1-\delta-\delta^2)(1/L)C$ and increases in his own punishment component $r_i(\cdot)$ by at least $(\delta^2/I)c_0$ (and by the economic property, does not decrease in his opponents' punishment components $r_{-i}(\cdot)$). The second term exceeds the first term by (20).

We conclude that for no conjecture is m_i a best response, contradicting our original assumption. This proves our second claim. □

While the basic construction of this proof follows Abreu and Matsushima (1992c), some complications arise in our robust formulation. The messages sent in the maximally revealing mechanism do not partition an agent's types. Rather, for each set of types that survives the iterated deletion of sets that can always be separated, there is a message that may be sent by all types in that set. So we say that message m_i^l is consistent with m_i^0 if message m_i^0 is one that might be sent by $m_i^0 \in S_i[\mathscr{M}^*](m_i^l)$.

The economic property can be weakened along the lines of Assumption 2 in Abreu and Matsushima (1992c). It would be enough to have the economic property hold for any type set profile Ψ in the inseparable type set Ξ^*, i.e. for each set profile $\Psi = (\Psi_i)_{i=1}^I \in \Xi^*$, there exists $(z_i)_{i=1}^I$ such that, for each i and $\theta \in \times_{i=1}^I \Psi_j$, $u_i(z_i,\theta) > u_i(\bar{y},\theta)$ and $u_j(\bar{y},\theta) \geq u_j(z_i,\theta)$ for all $j \neq i$.

6. Discussion

6.1 *Abreu–Matsushima measurability*

We establish in the preceding section that robust measurability, jointly with ex post incentive compatibility, is a necessary and sufficient condition for robust virtual implementation. Ex post incentive compatibility is equivalent to Bayesian incentive compatibility on the union of all type spaces (Bergemann and Morris 2005b). We now show

that robust measurability is equivalent to requiring that the notion of measurability originally suggested by Abreu and Matsushima (1992c) holds on the union of all type spaces.[8] To spell out the details of this equivalence result, we need a formal language for epistemic type spaces in the sense of Harsanyi (1967–68) and Mertens and Zamir (1985).

A *type space* is defined by $\mathcal{T} \triangleq (T_i, \widehat{\pi}_i, \widehat{\theta}_i)_{i=1}^I$, where each T_i is a countable set of types, the function $\widehat{\pi}_i : T_i \to \Delta(T_{-i})$ defines the beliefs that agent i assigns to other agents having types t_{-i}, and the function $\widehat{\theta}_i : T_i \to \Theta_i$ defines the agent i's payoff types. A type space is *finite* if each T_i is finite. We fix a type space $\mathcal{T}$ and write $\succeq_{t_i}^{\mathcal{T}}$ for the induced preferences of type t_i of agent i over type-contingent lotteries $\widetilde{y}_i : T_{-i} \to Y$. Thus

$$\widetilde{y}_i \succeq_{t_i} \widetilde{y}_i' \quad \text{if and only if} \quad \sum_{t_{-i} \in T_{-i}} \widehat{\pi}_i(t_{-i} \mid t_i) u_i(\widetilde{y}_i(t_{-i}), t) \geq \sum_{t_{-i} \in T_{-i}} \widehat{\pi}_i(t_{-i} \mid t_i) u_i(\widetilde{y}_i'(t_{-i}), t).$$

Fix a partition profile $\mathcal{H} = (\mathcal{H}_i)_{i=1}^I$, where each $\mathcal{H}_i$ is a partition of T_i. A function $\widetilde{y}_i : T_{-i} \to Y$ is $\mathcal{H}$-measurable if for all $j \neq i$

$$\{t_j, t_j'\} \subseteq H_j \in \mathcal{H}_j \Rightarrow \widetilde{y}(t_j, t_{-\{i,j\}}) = \widetilde{y}(t_j', t_{-\{i,j\}}).$$

Say that types t_i and t_i' are $(\mathcal{T}, \mathcal{H})$-distinguishable if there exists a $\mathcal{H}$-measurable $\widetilde{y}_i : T_{-i} \to Y$ such that

$$\widetilde{y}_i \succ_{t_i}^{\mathcal{T}} \bar{y} \quad \text{and} \quad \bar{y} \succ_{t_i'}^{\mathcal{T}} \widetilde{y}_i,$$

where we continue to denote by $\bar{y}$ the constant uniform lottery.

Now iteratively define a sequence of partitions $\mathcal{H}^k = (\mathcal{H}_i^k)_{i=1}^I$ by letting each $\mathcal{H}_i^0$ be the coarsest partition of the type set T_i, namely $\{T_i\}$, and letting each $\mathcal{H}_i^{k+1}$ consist of sets of types of agent i that are $(\mathcal{T}, \mathcal{H}^k)$-indistinguishable.

Let $\mathcal{H}^*$ be the limit of the sequence of partitions. We say that types t_i and t_i' are Abreu–Matsushima, or "AM", indistinguishable on type space $\mathcal{T}$, written $t_i \sim_{AM}^{\mathcal{T}} t_i'$, if t_i and t_i' are in the same element of the partition $\mathcal{H}_i^*$.

Proposition 3 (Equivalence).

(i) If θ_i and θ_i' are pairwise inseparable, then there exists a finite type space $\mathcal{T}$ and types $t_i, t_i' \in T_i$ such that (a) $\widehat{\theta}_i(t_i) = \theta_i$, (b) $\widehat{\theta}_i(t_i') = \theta_i'$, and (c) $t_i \sim_{AM}^{\mathcal{T}} t_i'$.

(ii) Conversely, if there exists a type space $\mathcal{T}$ (perhaps infinite but countable) and types $t_i, t_i' \in T_i$ such that (a) $\widehat{\theta}_i(t_i) = \theta_i$, (b) $\widehat{\theta}_i(t_i') = \theta_i'$, and (c) $t_i \sim_{AM}^{\mathcal{T}} t_i'$, then θ_i and θ_i' are pairwise inseparable.

The equivalence result of Proposition 3 suggests an alternative route to establishing the necessity result for robust implementation in Theorem 2: by the equivalence of robust measurability and AM measurability on the union of all type spaces, we could prove the necessity by an appeal to the arguments used in Abreu and Matsushima (1992c). By contrast, our sufficiency result (Theorem 3) cannot be established using the arguments

[8] We would like to thank an anonymous referee who suggested that we investigate the exact relationship between Abreu–Matsushima measurability and robust measurability.

and methods in Abreu and Matsushima (1992c): as the union of all type spaces is not a finite object, the arguments in Abreu and Matsushima (1992c)—which rely on the finiteness of the type space—cannot be applied.

6.2 *Interdependence and pairwise separability*

We illustrate the notions of pairwise and mutual inseparability in Section 3 in the context of a linear model of interdependent preferences for a single object:

$$v_i(\theta_i,\theta_{-i})=\theta_i+\gamma\sum_{j\neq i}\theta_j.$$

In this linear and symmetric model the parameter γ represents the level of interdependence in the preferences of the agents. We show that for $\gamma<1/(I-1)$, all payoff types of all agents are pairwise separable and suggest that pairwise separability requires not too much interdependence in the preferences.

We now establish the relationship between pairwise inseparability and moderate interdependence in a substantially more general environment. We assume that the utility function of each agent i is given by a convex combination of a private value utility function v_i and an interdependent utility function w_i over the general space of lotteries Y defined in Section 2. The private value utility function

$$v_i : Y\times\Theta_i\to\mathbb{R},$$

gives rise to distinct preferences for every θ_i:

$$\theta_i\neq\theta_i'\Rightarrow v_i(\cdot,\theta_i) \text{ is not an affine transformation of } v_i(\cdot,\theta_i').$$

The interdependent utility function

$$w_i : Y\times\Theta_{-i}\to\mathbb{R}$$

can depend in an arbitrary way on the type profile $\theta_{-i}\in\Theta_{-i}$ of all agents except agent i. For any $\gamma_i\in[0,1]$, let $u_i^{\gamma_i}$ be the utility function that puts weight $1-\gamma_i$ on the private value utility v_i and weight γ_i on the interdependent utility w_i:

$$u_i^{\gamma_i}(y,\theta)\triangleq(1-\gamma_i)v_i(y,\theta_i)+\gamma_i w_i(y,\theta_{-i}).$$

The interdependence in the preferences is now described by the vector of weights $\gamma=(\gamma_1,\ldots,\gamma_I)\in[0,1]^I$. For $\gamma=(0,\ldots,0)$ all payoff types of all agents are pairwise separable as, by assumption, the private utility function v_i gives rise to distinct preferences for all θ_i. Also, for $\gamma=(1,\ldots,1)$, we cannot separate any pair of types for any agent. In this case, the preferences of each agent are independent of his payoff type and therefore we cannot expect to separate the payoff types of agent i on the basis of his revealed preference. We parametrize the limit set Ξ^* that by Definition 2 describes the set of pairwise inseparable types, by the vector γ, or $\Xi^*(\gamma)$.

PROPOSITION 4 (Interdependence).

(i) The collection of sets $\Xi_i^*(\gamma)$ *satisfies* $\Xi_i^*(\mathbf{0}) = \{\{\theta_i^1\},\ldots,\{\theta_i^S\}\}$ *and* $\Xi_i^*(\mathbf{1}) = 2^{\Theta_i} \setminus \varnothing$ *for all* i.

(ii) If $\widehat{\gamma} \geq \gamma$, *then* $\Xi_i^*(\gamma) \subseteq \Xi_i^*(\widehat{\gamma})$.

The first part of the proposition determines the structure of the pairwise separable types with minimal and maximal interdependence. The second part establishes that the sets of pairs of types θ_i and θ_i' that are inseparable are weakly increasing in the interdependence parameter γ. In particular, it shows that the separability is monotone in the parameter of interdependence. We should emphasize that as the interdependence is represented by the vector $\gamma = (\gamma_1,\ldots,\gamma_I)$, the threshold for complete separability of all types and all agents itself is a multidimensional surface in the I-dimensional hypercube.

6.3 *Intermediate robustness notions*

The classic Bayesian implementation literature considers implementation on a fixed type space. We believe that this approach—as usually formulated—assumes too much common knowledge (among the agents and the planner) about the environment. In relaxing these common knowledge assumptions, we take an extreme approach: we maintain the assumption that there is common knowledge of the payoff structure of the environment (i.e., the set of possible payoff types of each agent and how each agent's utility function depends on the profile of payoff types) but do not restrict agents' beliefs and higher order beliefs about other agents' types.

In a recent paper, Artemov et al. (2008) consider what happens to robust virtual implementation results if one imposes some restrictions on agents' beliefs in the payoff environment. In particular, call a pair $(\theta_i,\lambda_i) \in \Theta_i \times \Delta(\Theta_{-i})$ a "pseudo-type" and suppose that we add the common knowledge that agent i's pseudo-type (θ_i,λ_i) belongs to a subset $T_i \subseteq \Theta_i \times \Delta(\Theta_{-i})$. When can a social choice function be virtually implemented on all type spaces where each agent i's pseudo-type belongs to T_i? Note that an agent's pseudo-type pins down his payoff type and belief about others' payoff types, but not his higher order beliefs. Thus this assumption is intermediate between the standard approach and our robustness approach. In the special case where $T_i = \Theta_i \times \Delta(\Theta_{-i})$, this setting becomes the setting of this paper. But if T_i is a strict subset of $\Theta_i \times \Delta(\Theta_{-i})$, the conditions for robust virtual implementation are weakened.

Now say that "pseudo-type diversity" is satisfied if

1. The set of beliefs consistent with a payoff type is a compact set, i.e., $\{\lambda_i \in \Delta(\Theta_{-i}) \mid (\theta_i,\lambda_i) \in T_i\}$ is a compact set for each i and $\theta_i \in \Theta_i$.

2. Two distinct payoff types cannot have the same preference over constant lotteries, i.e., $(\theta_i,\lambda_i),(\theta_i',\lambda_i') \in T_i$ and $\theta_i \neq \theta_i' \Rightarrow R_{\theta_i,\lambda_i} \neq R_{\theta_i',\lambda_i'}$.

Artemov et al. (2008) show that if pseudo-type diversity is satisfied, then robust virtual implementation is always possible if the appropriate incentive compatibility conditions are satisfied (their Theorem 1). The idea is that agents' payoff types can then be

identified by their preferences over constant lotteries and the Abreu and Matsushima (1992c)-style argument applied.[9]

To get a feel for the strength of the pseudo-type diversity condition, we can return to our leading example in Section 3. Recall that each $\Theta_i = [0,1]$ and $v_i = \theta_i + \gamma \mathbb{E}_i[\sum_{j \neq i} \theta_j]$ is a sufficient statistic for agent i's preferences. Now let $\Lambda_i \subseteq \Delta([0,1]^{I-1})$ be a compact set of beliefs over others' types that agent i may have (whatever his payoff type), so his set of possible pseudo-types is the product set $T_i = [0,1] \times \Lambda_i$. Now if $0 < \gamma \leq 1/(I-1)$, so there is not too much interdependence of preferences, pseudo-type diversity will be satisfied if and only if each Λ_i is a singleton.[10]

Artemov et al. (2008) also report the appropriate measurability condition required for robust virtual implementation if the pseudo-type diversity condition fails (their Definition 12 and Theorem 2). This is naturally intermediate between Abreu–Matsushima measurability and our robust measurability condition. We can illustrate this also with our example. Suppose that the probability that agent i assigns to any subset of other agents' payoff types is always at least $1-\delta$ times the probability of that event under a uniform prior, so that

$$\Lambda_i = \left\{ \lambda_i \in \Delta(\Theta_{-i}) \mid \lambda_i(E) \geq (1-\delta) \int_{\theta_{-i} \in E} d\theta_{-i}, \ \forall \text{ measurable } E \subseteq [0,1]^{I-1} \right\}$$

and $T_i = \Theta_i \times \Lambda_i$.

Now suppose that agent i's payoff type is in Ψ_i and he knows that other agents' payoff types are in Ψ_{-i}. If agent i's beliefs are restricted to belong to Λ_i, when does there exist a pair of payoff types in Ψ_i who could not have the same expected valuation of the object? Only if

$$\max \Psi_i + \gamma \sum_{j \neq i} ((1-\delta)\tfrac{1}{2} + \delta \min \Psi_j) > \min \Psi_i + \gamma \sum_{j \neq i} ((1-\delta)\tfrac{1}{2} + \delta \max \Psi_j).$$

Thus Ψ_{-i} "δ-separates" Ψ_i if and only if

$$\max \Psi_i - \min \Psi_i \leq \gamma \delta \sum_{j \neq i} (\max \Psi_j - \min \Psi_j).$$

Now the argument of Section 3 can be adapted to show that if $\gamma\delta < 1/(I-1)$, all pairs of distinct payoff types are strategically distinguishable from each other (under δ belief

[9]The version of "pseudo-type diversity" that we report is sufficient to implement the social choice functions depending just on payoff types that we study in this paper. Artemov et al. (2008) assume a slightly stronger version of pseudo-type diversity: they assume that each T_i is finite and that distinct pseudo-types have distinct preferences over constant lotteries even if they correspond to the same payoff type, i.e., $(\theta_i, \lambda_i), (\theta_i', \lambda_i') \in T_i$ and $(\theta_i, \lambda_i) \neq (\theta_i', \lambda_i') \Rightarrow R_{\theta_i, \lambda_i} \neq R_{\theta_i', \lambda_i'}$. This allows them to implement richer social choice functions that treat types with the same payoff types (but different beliefs over others' payoff types) differently.

[10]This example has a continuum of payoff types, so does not fit our formal framework. But we could make the same point with a finite grid of payoff types.

restrictions) and thus incentive compatibility is sufficient for robust virtual implementation. And if $\gamma\delta > 1/(I-1)$, all pairs of payoff types are strategically indistinguishable from each other (under δ belief restrictions) and robust virtual implementation is impossible for any (non-constant) social choice function.

6.4 *Rationalizability and all equilibria on all type spaces*

Our analysis takes as given the solution concept of incomplete information rationalizability for our environment. Thus we assume that if the agents' true payoff type profile was $\theta = (\theta_1, \ldots, \theta_I)$, they might send any message profile

$$m \triangleq (m_1, \ldots, m_I) \in \underset{i=1}{\overset{I}{\times}} S_i[\mathscr{M}](\theta_i) \triangleq S[\mathscr{M}](\theta).$$

Our motivation for employing this solution concept is that we do not want to make any assumption about agents' beliefs and higher order beliefs about other agents' payoff types. In fact, suppose one constructed a "type space" $\mathscr{T}$ specifying for each agent a set of possible epistemic types, and, for each epistemic type, a description of his (known) payoff type and his beliefs about others' epistemic types. By standard universal type space arguments, we can incorporate any beliefs and higher order beliefs about others' payoff types in such a type space. Now the type space $\mathscr{T}$ and a mechanism $\mathscr{M}$ together define a standard incomplete information game. The set of messages that can be sent by *any* type of agent i with payoff type θ_i in *any* Bayesian Nash equilibrium of the game $(\mathscr{T}, \mathscr{M})$ for *any* type space $\mathscr{T}$ is equal to $S_i[\mathscr{M}](\theta_i)$. This result is the straightforward incomplete information extension of the classic epistemic foundations result of Brandenburger and Dekel (1987), showing that the set of actions that can be played in the subjective correlated equilibria of a complete information game equals the set of actions that survive iterated deletion of strictly dominated actions in that game. Battigalli and Siniscalchi (2003) report the incomplete information version of this result as Propositions 4.2 and 4.3. For completeness, we formally state and prove this result in the appendix of the working paper version (Bergemann and Morris 2007).

This observation means that the gap between the solution concepts of pure strategy Bayesian Nash equilibrium (Serrano and Vohra 2001, 2005) and iterated deletion of (interim) strictly dominated strategies (Abreu and Matsushima 1992c) in incomplete information virtual implementation disappears in our robust approach. We consider this to be an attraction of our approach. The intuition is that the extra bite obtained by the assumption of equilibrium is lost without complementary strong assumptions on beliefs and higher order beliefs for the implementation problem.

6.5 *Iterated deletion of weakly dominated strategies*

Our incomplete information rationalizability solution concept is equivalent to iterated deletion of strictly dominated strategies. What happens if we look at iterated deletion of

weakly dominated strategies instead? In other words, we let $W_i^0[\mathscr{M}](\theta_i) = M_i$,

$$W_i^{k+1}[\mathscr{M}](\theta_i) = \left\{ m_i \in W_i^k[\mathscr{M}](\theta_i) \;\middle|\; \begin{array}{l} \exists\, \mu_i \in \Delta_{++}\{(\theta_{-i}, m_{-i}) \mid m_{-i} \in W_{-i}^k[\mathscr{M}](\theta_{-i})\} \text{ s.t.} \\ \\ m_i \in \arg\max_{m_i'} \sum_{\theta_{-i}, m_{-i}} \mu_i(\theta_{-i}, m_{-i}) u_i(g(m_i', m_{-i}), (\theta_i, \theta_{-i})) \end{array} \right\}$$

and

$$W_i[\mathscr{M}](\theta_i) = \bigcap_{k \geq 0} W_i^k[\mathscr{M}](\theta_i).$$

It is easy to see that our "negative" results go through unchanged. If two types are pairwise inseparable ($\theta_i \sim \theta_i'$) then the argument of Proposition 1—unchanged—implies that they have iteratively weakly undominated actions in common in every mechanism, or

$$W_i[\mathscr{M}](\theta_i) \cap W_i[\mathscr{M}](\theta_i') \neq \varnothing \text{ for all } \mathscr{M}.$$

Thus robust measurability is a necessary condition for implementation (virtual or exact) of any social choice function in iterated deletion of weakly dominated strategies in a finite (or compact) mechanism: the argument of Theorem 2 goes through unchanged in this case.

Abreu and Matsushima (1994) show that their argument for *virtual* complete information implementation in iterated deletion of *strictly* dominated strategies can be adapted to show the possibility of *exact* complete information implementation in iterated deletion of *weakly* dominated strategies, with some extra restrictions on the environment. It is a reasonable conjecture that this extension can be adapted to the standard incomplete information implementation setting of Abreu and Matsushima (1992c) and our robust incomplete information setting. However, we have not attempted this extension.

Chung and Ely (2001) show that in an auction environment with interdependent valuations as in Section 3, the efficient outcome can be implemented in the direct mechanism under iterated deletion of weakly dominated strategies (i.e., the solution concept described above) under the assumption that $\gamma < 1/(I-1)$. Our results supply a strong converse: if $\gamma \geq 1/(I-1)$, it is not possible to implement (exactly or virtually) any nontrivial social choice function in iterated deletion of weakly dominated strategies in any finite (or compact) mechanism, direct or indirect.[11]

6.6 *Implementation in a direct mechanism*

We restrict attention in this paper to finite mechanisms. Thus the mechanisms here do not include any of the pathological features of "integer games" that play an important role in the full implementation literature and have been much criticized (see, e.g.,

[11] Our results are stated for a lottery space over finite outcomes, but the extension to any compact space and compact mechanisms is straightforward.

Jackson 1992). Nonetheless, the mechanisms in this paper are complex. The canonical mechanism for robust virtual implementation inherits the complexity of the mechanism of Abreu and Matsushima (1992c), on which it builds. Our maximally revealing mechanism generating strategic distinguishability is no simpler. While the mechanisms are theoretically kosher, it has been argued that their complexity and the logic of the iteration deletion in the mechanism might make them hard to use in practise. For example, Glazer and Rosenthal (1992) have made this argument about the mechanism used by Abreu and Matsushima (1992b) for complete information virtual implementation (see Abreu and Matsushima 1992a for a response and Sefton and Yavas 1996 for later experiments inspired by the mechanism).

By requiring robustness to agents' beliefs and higher order beliefs, we reduce the amount of common knowledge about the environment that can be used by the planner in designing a mechanism. This makes it harder to achieve positive results (and our robust measurability condition is rather strong in applications). But one motivation for studying robust implementation is that we hope that robustness considerations will endogenously lead to simpler mechanisms when positive results can be achieved. By adapting results from our earlier work on exact robust implementation in direct mechanisms (Bergemann and Morris forthcoming), we can report that, in at least one broad class of economic environments of interest, whenever robust virtual implementation is possible according to Corollary 1, it is possible in a direct mechanism where agents simply report their payoff types. We say that preferences satisfy *aggregator single crossing* (ASC) if each agent i's preferences at type profile θ belong to a single crossing class parameterized by $h_i(\theta)$, where $h_i : \Theta \to \mathbb{R}$ is a monotonic aggregator. Bergemann and Morris (forthcoming) established that exact robust implementation by a compact mechanism is possible if and only if the social choice function satisfies strict ex post incentive compatibility and a *contraction property* on the aggregator functions $h = (h_1, \ldots, h_I)$. In the appendix of the working paper version, we show that under the ASC assumption, robust measurability is always satisfied under the contraction property.

6.7 *Exact implementation and integer games*

The first papers on incomplete information implementation focus on exact implementation. Postlewaite and Schmeidler (1986) and Jackson (1991) identify a Bayesian monotonicity condition that (together with Bayesian incentive compatibility) is necessary and (under weak economic conditions) sufficient for exact implementation in Bayesian Nash equilibrium. Bergemann and Morris (2005a) provide a robust analogue of this result, showing that ex post incentive compatibility and a *robust monotonicity* condition are necessary and—under weak economic conditions—sufficient for exact robust implementation. All these papers follow a tradition in the implementation literature of allowing very badly behaved mechanisms, like integer games, in proving their general results. In this paper, we follow Abreu and Matsushima (1992c) in restricting attention to finite—and thus well-behaved—mechanisms. We briefly discuss the relation between these results in this section; a more complete and formal discussion in contained in the appendix of the working paper version (Bergemann and Morris 2007).

Robust measurability and robust monotonicity turn out to be equivalent in the important class of aggregator single crossing preferences. However, in general, one can show by example that robust measurability neither implies nor is implied by robust monotonicity. Thus requiring only virtual implementation is sometimes a strict relaxation, and allowing badly-behaved mechanisms is sometimes a strict relaxation. We do not have a characterization of when exact robust implementation by a well behaved mechanism is possible (just as analogous characterizations do not exist for complete information and classical Bayesian implementation). We know only that robust measurability, robust monotonicity, and strict ex post incentive compatibility are all necessary.

We restrict attention in our analysis to social choice functions rather than social choice correspondences. Bergemann and Morris (2005b) consider the problem of *partially* robustly implementing a social choice correspondence, i.e., ensuring that whatever players' beliefs and higher order beliefs about others' types, there is *an* equilibrium leading to outcomes contained in the social choice correspondence. In the special case where the social choice correspondence is a function (and more generally in a class of separable environments), this is possible only if the function (or a selection from the correspondence in separable environments) is ex post incentive compatible. But in the general case, we do not have a satisfactory characterization of when partial robust implementation is possible. For this reason, we have not attempted a characterization of (full) robust implementation of social choice correspondences.

Appendix

This appendix contains omitted proofs from the main body of the paper.

Proof of Lemma 1. Suppose that

$$\sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(\bar{y},(\theta_i,\theta_{-i})) \geq \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(x,(\theta_i,\theta_{-i})) \tag{22}$$

for all $x \in X$. If

$$\sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(\bar{y},(\theta_i,\theta_{-i})) > \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(x',(\theta_i,\theta_{-i}))$$

for some $x' \in X$, we could conclude that

$$\begin{aligned}\sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(\bar{y},(\theta_i,\theta_{-i})) &> \frac{1}{N}\sum_{x\in X}\sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(x,(\theta_i,\theta_{-i})) \\ &= \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(\bar{y},(\theta_i,\theta_{-i})),\end{aligned}$$

a contradiction. So (22) implies

$$\sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(\bar{y},(\theta_i,\theta_{-i})) = \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(x,(\theta_i,\theta_{-i})) \tag{23}$$

for all $x \in X$. But (23) implies that R_{θ_i,λ_i} is indifferent between all pure outcomes and thus all lotteries. This contradicts Assumption 1 of no-complete-indifference. We conclude that the no-complete-indifference assumption implies that (22) fails for all i, i.e., that for all i, $\theta_i \in \Theta_i$ and $\lambda_i \in \Delta(\Theta_{-i})$, there exists $x \in X$ such that

$$\sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(x,(\theta_i,\theta_{-i})) > \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})u_i(\bar{y},(\theta_i,\theta_{-i})).$$

Equivalently, for all i, $\theta_i \in \Theta_i$ and $\lambda_i \in \Delta(\Theta_{-i})$,

$$\max_{x\in X} \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})[u_i(x,(\theta_i,\theta_{-i})) - u_i(\bar{y},(\theta_i,\theta_{-i}))] > 0.$$

Now, note that for each $x \in X$ the function

$$\sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i})[u_i(x,(\theta_i,\theta_{-i})) - u_i(\bar{y},(\theta_i,\theta_{-i}))]$$

is continuous in λ (in the standard topology). The conclusion follows from the compactness (in the standard topology) of $\Delta(\Theta_{-i})$ and continuity of the maximum operator. □

Proof of Lemma 4. Fix any $m_i \notin S_i[\mathcal{M}](\theta_i)$. Then there exists k such that $m_i \in S_i^k[\mathcal{M}](\theta_i)$ and $m_i \notin S_i^{k+1}[\mathcal{M}](\theta_i)$. Consider

$$\Delta_i^k = \{\mu_i \in \Delta(\Theta_{-i} \times M_{-i}) \mid \mu_i(\theta_{-i} \times m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^k[\mathcal{M}](\theta_{-i}) \text{ for each } j \neq i\}.$$

For all $\mu_i \in \Delta_i^k$, there exists $\bar{m}_i$ such that

$$\sum_{\theta_{-i},m_{-i}} \mu_i(\theta_{-i},m_{-i})u_i(g(\bar{m}_i,m_{-i}),(\theta_i,\theta_{-i})) > \sum_{\theta_{-i},m_{-i}} \mu_i(\theta_{-i},m_{-i})u_i(g(m_i,m_{-i}),(\theta_i,\theta_{-i})).$$

By the compactness of Δ_i^k, there exists $\bar{\varepsilon}_i(m_i) > 0$ such that for all $\mu_i \in \Delta_i^k$ there exists $\bar{m}_i$ such that

$$\sum_{\theta_{-i},m_{-i}} \mu_i(\theta_{-i},m_{-i})u_i(g(\bar{m}_i,m_{-i}),(\theta_i,\theta_{-i})) > \sum_{\theta_{-i},m_{-i}} \mu_i(\theta_{-i},m_{-i})u_i(g(m_i,m_{-i}),(\theta_i,\theta_{-i})) + \bar{\varepsilon}_i(m_i).$$

Now let

$$\eta_{\mathcal{M}} = \min_{i,\ \theta_i \text{ and } m_i\notin S_i[\mathcal{M}](\theta_i)} \bar{\varepsilon}_i(m_i),$$

which establishes the desired bound. □

Proof of Lemma 5. Suppose $\pi^+ C \leq \pi^0 \eta_{\mathcal{M}^0}$. We argue, by induction on k, that

$$(m_i^0, m_i^1) \in S_i^k[\mathcal{M}](\theta_i) \Rightarrow m_i^0 \in S_i^k[\mathcal{M}^0](\theta_i)$$

for all $k \geq 0$. This is true by definition for $k = 0$. Suppose that it is true for k. Now suppose that $m_i^0 \notin S_i^{k+1}[\mathscr{M}^0](\theta_i)$ but $(m_i^0, m_i^1) \in S_i^{k+1}[\mathscr{M}](\theta_i)$ and so $(m_i^0, m_i^1) \in S_i^k[\mathscr{M}](\theta_i)$ and, by the inductive hypothesis, $m_i^0 \in S_i^k[\mathscr{M}^0](\theta_i)$. Now fix any $\mu_i \in \Delta(\Theta_{-i} \times M_{-i})$ satisfying

$$\mu_i(\theta_{-i}, (m_j^0, m_j^1)_{j \neq i}) > 0 \Rightarrow (m_j^0, m_j^1)_{j \neq i} \in S_{-i}^k[\mathscr{M}](\theta_{-i}) \Rightarrow m_{-i}^0 \in S_{-i}^k[\mathscr{M}^0](\theta_{-i}).$$

Let

$$\bar{\mu}_i(\theta_{-i}, m_{-i}^0) = \sum_{(m_j^1)_{j \neq i} \in M_{-i}^1} \mu_i(\theta_{-i}, (m_j^0, m_j^1)_{j \neq i}).$$

By Lemma 4, there exists $\bar{m}_i^0$ such that

$$\sum_{\theta_{-i}, m_{-i}^0} \bar{\mu}_i(\theta_{-i}, m_{-i}^0)[u_i(g^0(\bar{m}_i^0, m_{-i}^0), (\theta_i, \theta_{-i})) - u_i(g^0(m_i^0, m_{-i}^0), (\theta_i, \theta_{-i}))] > \eta_{\mathscr{M}^0}.$$

Thus

$$\sum_{\theta_{-i}, m_{-i}} \mu_i(\theta_{-i}, m_{-i})[u_i(g((\bar{m}_i^0, m_i^1), m_{-i}), (\theta_i, \theta_{-i})) - u_i(g((m_i^0, m_i^1), m_{-i}), (\theta_i, \theta_{-i}))] > \pi^0 \eta_{\mathscr{M}^0} - \pi^+ C \geq 0.$$

This contradicts our premise that $(m_i^0, m_i^1) \in S_i^{k+1}[\mathscr{M}](\theta_i)$, and we conclude that $(m_i^0, m_i^1) \in S_i^{k+1}[\mathscr{M}](\theta_i) \Rightarrow m_i^0 \in S_i^{k+1}[\mathscr{M}^0](\theta_i)$. □

The canonical mechanism asks each agent to make a series of binary choices between the central lottery $\bar{y}$ and a specific lottery y from the test set. If the test set is to be successful in eliciting the private information from agent i, then the test set should contain a sufficient number of allocations such that for every type θ_i and every belief λ_i of agent i there exists some allocation y that is strictly preferred to the central lottery $\bar{y}$.

Lemma 9 (Duality). *The type set profile Ψ_{-i} separates Ψ_i if and only if there exists $\tilde{y} : \Psi_i \to Y$ such that*

$$\sum_{\theta_i \in \Psi_i} (\tilde{y}(\theta_i) - \bar{y}) = 0 \tag{24}$$

and

$$\tilde{y}(\theta_i) \, P_{\theta_i, \lambda_i} \, \bar{y} \tag{25}$$

for all $\theta_i \in \Psi_i$ and all $\lambda_i \in \Delta(\Psi_{-i})$.

This result says that for each $\theta_i \in \Psi_i$, we can identify a direction in the lottery space, $\tilde{y}(\theta_i) - \bar{y}$, that agent i likes whatever his beliefs about Ψ_{-i}, such that the sum of those changes add up to zero. The lemma follows from the following duality result in Samet (1998).

PROPOSITION 5 (Samet 1998). *Let $V_1,\dots,V_S$ be closed, convex, subsets of the N-dimensional simplex Δ^N. These sets have an empty intersection if and only if there exist $z_1,\dots,z_S\in\mathbb{R}^N$ such that*

$$\sum_{s=1}^{S} z_s = 0$$

and

$$v\cdot z_s > 0 \text{ for each } s=1,\dots,S \text{ and } v\in V_s.$$

This result is introduced in Samet (1998) to provide a simple proof of the observation that asymmetrically informed agents trade against each other if and only if they do not share a common prior, i.e., their posterior beliefs cannot be derived by updating a common prior.[12] Suppose that there are N states and S agents. Each agent s observes one of a collection of signals about the true state. Each signal leads him to have a posterior $v\in\Delta^N$ over the states. Let V_s be the convex hull of his set of possible posteriors. Notice that V_s represents the set of prior beliefs he might have held over the state space before observing his signal. Thus posterior beliefs are consistent with a common prior if and only if the intersection of the V_s sets is non-empty. Now consider a multilateral bet specifying that if state n is realized, agent s receives payment z_{sn} where the total payments sum to zero:

$$\sum_{s=1}^{S} z_{sn} = 0 \text{ for all } n.$$

Writing $z_s \triangleq (z_{sn})_{n=1}^N$, we then have

$$\sum_{s=1}^{S} z_s = 0.$$

There exists such a bet where every agent has a strictly positive expected value from accepting the bet conditional on every signal if $v\cdot z_s>0$ for each $s=1,\dots,S$ and $v\in V_s$.

PROOF OF LEMMA 9. By definition, the type set profile Ψ_{-i} separates Ψ_i if, for every $R\in\mathscr{R}$, there exists $\theta_i\in\Psi_i$ such that $R_{\theta_i,\lambda_i}\neq R$ for every $\lambda_i\in\Delta(\Psi_{-i})$. Write

$$\begin{aligned} X &= \{x_1,\dots,x_n,\dots,x_N\}\\ \Theta_i &= \{\theta_i^1,\dots,\theta_i^s,\dots,\theta_i^S\}\\ \Theta_{-i} &= \{\theta_{-i}^1,\dots,\theta_{-i}^w,\dots,\theta_{-i}^W\}, \end{aligned}$$

with $W=S^{I-1}$. The vector

$$v_{sw} = (u_i(x_n,(\theta_i^s,\theta_{-i}^w)))_{n=1}^N$$

is an element of $\mathbb{R}^N$. Without loss of generality (since expected utility preferences can be represented by any affine transformation), we can assume that each v_{sw} is an element

[12] This converse to the no trade theorem was originally proved by Morris (1994), by a more indirect duality argument.

of the N-dimensional simplex Δ^N. Now $(v_{sw})_{w=1}^W$ is a collection of W elements of Δ^N, and the set of preferences

$$\{R_{\theta_i^s,\lambda_i} : \lambda_i \in \Delta(\Psi_{-i})\}$$

is represented by the convex hull of $(v_{sw})_{w=1}^W$, which we write as

$$V_s = \text{conv}\left((v_{sw})_{w=1}^W\right) \subseteq \Delta^N.$$

Thus Ψ_{-i} separates Ψ_i exactly if

$$\bigcap_{s=1}^S V_s = \varnothing.$$

By Proposition 5, this is true if and only if there exist $z_1, \ldots, z_S \in \mathbb{R}^N$ such that

$$\sum_{s=1}^S z_s = 0 \tag{26}$$

and

$$v \cdot z_s > 0 \tag{27}$$

for all s and all $v \in V_s$. But if $(z_s)_{s=1}^S$ satisfy (26) and (27), we may choose $\varepsilon > 0$ sufficiently small such that $\widetilde{y}(\theta_i^s) = \bar{y} + \varepsilon z_s \in Y$ for each s, and we have established (24) and (25).

Conversely, if (24) and (25) hold and we set $z_s = \widetilde{y}(\theta_i^s) - \bar{y}$ for $s = 1, \ldots, S$, then $(z_s)_{s=1}^S$ satisfy (26) and (27). □

We now use Lemma 9 to show how, if Ψ_{-i} separates Ψ_i, we can construct a *finite* set of lotteries $\widetilde{Y}_i(\Psi_i, \Psi_{-i}) \subseteq Y$ such that knowing that agent i knows that his opponent's type is in Ψ_{-i} and knowing his preferences on $\widetilde{Y}_i(\Psi_i, \Psi_{-i})$ is always enough to rule out at least one type in Ψ_i for agent i.

LEMMA 10. *If Ψ_{-i} separates Ψ_i, then there exists a finite set $\widetilde{Y}_i(\Psi_i, \Psi_{-i}) \subseteq Y$ such that for each $\theta_i \in \Psi_i$ and $\lambda_i \in \Delta(\Psi_{-i})$, there exists $y \in \widetilde{Y}_i(\Psi_i, \Psi_{-i})$ such that*

$$\bar{y} \; P_{\theta_i,\lambda_i} \; y \tag{28}$$

and for some $\theta_i' \in \Psi_i$,

$$y \; P_{\theta_i',\lambda_i'} \; \bar{y} \tag{29}$$

for all $\lambda_i' \in \Delta(\Psi_{-i})$.

PROOF. By Lemma 9, there exists $\widetilde{y} : \Psi_i \to Y$ such that

$$\sum_{\theta_i \in \Psi_i} (\widetilde{y}(\theta_i) - \bar{y}) = 0$$

and

$$\widetilde{y}(\theta_i) \; P_{\theta_i,\lambda_i} \; \bar{y} \text{ for all } \theta_i \in \Psi_i \text{ and } \lambda_i \in \Delta(\Psi_{-i}).$$

Let $\widetilde{Y}_i(\Psi_i,\Psi_{-i}) = \{\widetilde{y}(\theta_i)\}_{\theta_i\in\Psi_i}$. Fix $\theta_i \in \Psi_i$ and $\lambda_i \in \Delta(\Psi_{-i})$. Write $\widetilde{Y}_i(\Psi_i,\Psi_{-i}) = \{y^1,\dots,y^K\}$, with $y^1 = \widetilde{y}(\theta_i)$. Let $\bar{y}^0 = \bar{y}$ and

$$\bar{y}^l = \bar{y} + \varepsilon\sum_{\kappa=1}^{l}(y^\kappa - \bar{y}),$$

with $\varepsilon > 0$ chosen sufficiently small such that $\bar{y}^l \in Y$ for all $l = 1,\dots,K$. We know $\bar{y}^1\ P_{\theta_i,\lambda_i}\ \bar{y}^0$. Suppose $\bar{y}^{l+1}\ R_{\theta_i,\lambda_i}\ \bar{y}^l$ for all $l = 1,\dots,K-1$. By transitivity, this implies that $\bar{y}^K\ P_{\theta_i,\lambda_i}\ \bar{y}^0$. But $\bar{y}^K = \bar{y}^0$, so we have a contradiction. We conclude that, for some $l = 1,\dots,K-1$, $\bar{y}^l\ P_{\theta_i,\lambda_i}\ \bar{y}^{l+1}$. This implies that there exists θ_i' such that $\bar{y}\ P_{\theta_i,\lambda_i}\ y(\theta_i')$. Since

$$y(\theta_i')\ P_{\theta_i',\lambda_i'}\ \bar{y} \text{ for all } \lambda_i' \in \Delta(\Psi_{-i}),$$

the inequalities (28) and (29) are established. □

Now we construct a large enough finite set of lotteries (the "test set") such that knowing just an agent's most preferred outcome on the test set always reveals enough information about his preferences to separate out a type, if it is possible to do so.

The proof of Lemma 7 is constructive. We first construct a set $\widetilde{Y}$ consisting of the degenerate lotteries X and the sets $\widetilde{Y}_i(\Psi_i,\Psi_{-i})$ constructed in Lemma 10, for every triple (i,Ψ_i,Ψ_{-i}) with Ψ_{-i} separating Ψ_i. Knowing an agent's ranking of each element of $\widetilde{Y}$ relative to the central lottery $\bar{y}$ reveals all the information we need to extract. In order to extract this information in a single choice, we let the agent pick $f : \widetilde{Y} \to \{0,1\}$. For each $y \in \widetilde{Y}$, y is chosen with probability $1/\widetilde{Y}$ if $f(y) = 1$, otherwise the central lottery $\bar{y}$ is chosen. We let Y^* be the set of all such lotteries. Now observing an agent's most preferred outcome in Y^* reveals his binary preference between $\bar{y}$ and each element of $\widetilde{Y}$. Since $\widetilde{Y}$ contains each $\widetilde{Y}_i(\Psi_i,\Psi_{-i})$, this ensures part (ii). Since $\widetilde{Y}$ contains all the lotteries that put probability 1 on each pure outcome, Assumption 1 (no-complete-indifference) implies that, for each θ_i and λ_i, there exist $y,y' \in \widetilde{Y}$ such that $y\ P_{\theta_i,\lambda_i}\ y'$ and thus $y' \notin B_i^{Y^*}(\theta_i,\lambda_i)$. This proves part (i).

Proof of Lemma 7. Let

$$\widetilde{Y} = X \cup \bigcup_{\{(i,\Psi_i,\Psi_{-i})|\Psi_{-i} \text{ separates } \Psi_i\}} \widetilde{Y}_i(\Psi_i,\Psi_{-i}).$$

Now for any $f : \widetilde{Y} \to \{0,1\}$, let y_f be the lottery obtained by picking an element $y \in \widetilde{Y}$ with uniform probability and then choosing lottery y if $f(y) = 1$ and $\bar{y}$ if $f(y) = 0$. Thus we define

$$y_f \equiv \bar{y} + \frac{1}{\#\widetilde{Y}}\sum_{y\in\widetilde{Y}} f(y)(y - \bar{y}).$$

Let Y^* be the set of such lotteries, i.e.,

$$Y^* = \{y \in Y \mid \exists f : \widetilde{Y} \to \{0,1\} \text{ such that } y = y_f\}.$$

To prove part (i) of the lemma, fix any $\theta_i \in \Theta_i$ and $\lambda_i \in \Delta(\Theta_{-i})$. By Lemma 1, there exists $x \in X \subseteq \widetilde{Y}$ such that $x \; P_{\theta_i,\lambda_i} \; \bar{y}$; now let $f^0(y) = 0$, for all $y \in \widetilde{Y}$, and

$$f^*(y) = \begin{cases} 0 & \text{if } y \neq x \\ 1 & \text{if } y = x. \end{cases}$$

So we can write

$$y_{f^0} = \bar{y}, y_{f^*} = \bar{y} + \frac{1}{\#\widetilde{Y}}(x - \bar{y})$$

and so $y_{f^0} \notin B_i^{Y^*}(\theta_i, \lambda_i)$.

To prove part (ii) of the lemma, suppose that Ψ_{-i} separates Ψ_i. Fix $\theta_i \in \Psi_i$ and $\lambda_i \in \Delta(\Psi_{-i})$. By Lemma 10, there exists $y \in \widetilde{Y}_i(\Psi_i, \Psi_{-i})$ and $\theta_i' \in \Psi_i$ such that $\bar{y} \; P_{\theta_i,\lambda_i} \; y$ and $y \; P_{\theta_i',\lambda_i'} \; \bar{y}$ for all $\lambda_i' \in \Delta(\Psi_{-i})$. So

$$y_f \in B_i^{Y^*}(\theta_i, \lambda_i) \Rightarrow f(y) = 0,$$

while

$$y_f \in B_i^{Y^*}(\theta_i', \Psi_i) \Rightarrow f(y) = 1,$$

and so

$$B_i^{Y^*}(\theta_i, \lambda_i) \cap B_i^{Y^*}(\theta_i', \Psi_i) = \varnothing,$$

which establishes the result. □

Proof of Proposition 2. Consider types θ_i and θ_i' such that $\theta_i \nsim \theta_i'$. Then by the definition of pairwise inseparability, $\{\theta_i, \theta_i'\} \in \Xi_i^*$. By the construction of the inseparable sets Ξ_i^k, it follows that there is a finite stage $\bar{k}$ such that $\{\theta_i, \theta_i'\} \in \Xi_i^{\bar{k}}$ but

$$\{\theta_i, \theta_i'\} \notin \Xi_i^{\bar{k}+1}. \tag{30}$$

By Lemma 7, for all i, k and m_i^k we have

$$\bar{\Theta}_i^k(m_i^k) \in \Xi_i^k, \tag{31}$$

and by Lemma 6, for each k there exists $\bar{\varepsilon} > 0$ such that

$$\{\theta_i \in \Theta_i \mid m_i^k \in S_i[\mathcal{M}_\varepsilon^k](\theta_i)\} \subseteq \bar{\Theta}_i^k(m_i^k), \tag{32}$$

for all $\varepsilon < \bar{\varepsilon}$ and $m_i^k \in M_i^k$. Now since Ξ^* is established in a finite number of stages, it follows that by the choosing k sufficiently large and ε sufficiently small, we obtain an augmented mechanism $\mathcal{M}_\varepsilon^K = \mathcal{M}^*$such that if $\theta_i \nsim \theta_i'$, then from the exclusion (30) and the inclusions (31) and (32), it follows that $S_i^{\mathcal{M}^*}(\theta_i) \cap S_i^{\mathcal{M}^*}(\theta_i') = \varnothing$, which establishes the result. □

Proof of Proposition 3. (i) Fix mutually inseparable $\Xi = (\Xi_i)_{i=1}^I$. We use properties of Ξ to construct a type space $\mathcal{T}$. For each $\Psi_i \in \Xi_i$, there exists $\Psi_{-i}^{\Psi_i} \in \Xi_{-i}$ such that $\Psi_{-i}^{\Psi_i}$ does not separate Ψ_i. Recall that "$\Psi_{-i}^{\Psi_i}$ does not separate Ψ_i" means that there exists a

preference relation R_i over uncontingent lotteries Y such that for each $\theta_i \in \Psi_i$, there exists $\lambda_i^{\theta_i,\Psi_i} \in \Delta(\Psi_{-i}^{\Psi_i})$ such that $R_{\theta_i,\lambda_i^{\theta_i,\Psi_i}} = R_i$. Now, for each i, let

$$T_i \triangleq \{(\theta_i,\Psi_i) \in \Theta_i \times \Xi_i \mid \theta_i \in \Psi_i\}, \tag{33}$$

with

$$\widehat{\pi}_i((\theta_j,\Psi_j)_{j\neq i} \mid (\theta_i,\Psi_i)) \triangleq \begin{cases} \lambda_i^{\theta_i,\Psi_i}(\theta_{-i}) & \text{if } \Psi_{-i} = \Psi_{-i}^{\Psi_i} \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

and

$$\widehat{\theta}_i(\theta_i,\Psi_i) \triangleq \theta_i. \tag{35}$$

Now consider the partition $\mathscr{H}_i$ of the type set $\mathscr{T}_i$, as defined through (33)–(35), that is generated by the equivalence relation $(\theta_i,\Psi_i) \sim (\theta_i',\Psi_i')$ if $\Psi_i = \Psi_i'$. By construction, each (θ_i,Ψ_i) and (θ_i',Ψ_i) are $(\mathscr{T},\mathscr{H})$-indistinguishable. To see this, observe that since $\theta_i, \theta_i' \in \Psi_i$, there exists a common Ψ_{-i}, namely $\Psi_{-i}^{\Psi_i}$ such that $\lambda_i^{\theta_i,\Psi_i}(\Psi_{-i}^{\Psi_i}) = \lambda_i^{\theta_i',\Psi_i}(\Psi_{-i}^{\Psi_i}) = 1$. Now, as the type contingent lottery $\tilde{y}_i$ has to be $\mathscr{H}$-measurable, it follow in particular that it has to be constant on $\Psi_{-i}^{\Psi_i}$ and hence is an uncontingent lottery on $\Psi_{-i}^{\Psi_i}$. But Lemma 3 shows that if any payoff types θ_i and θ_i' are pairwise inseparable, then there exists a mutually inseparable $\Xi = (\Xi_i)_{i=1}^I$ and $\Psi_k \in \Xi_k$ with $\{\theta_k, \theta_k'\} \subseteq \Psi_k$.

(ii) For the other direction, fix a type space $\mathscr{T}$. Write $\mathscr{H}^*$ for the limit of the sequence of partitions defined above and let $\sim_{AM}^{\mathscr{T}}$ be the corresponding equivalence relation. Write $H_i^*(t_i) = \{t_i' | t_i' \sim_{AM}^{\mathscr{T}} t_i\}$ and let

$$\Xi_i \triangleq \{\Psi_i \in 2^{\Theta_i} \setminus \varnothing \mid \exists t_i \in T_i \text{ such that } \Psi_i = \{\theta_i \mid \exists t_i' \in H_i^*(t_i) \text{ with } \widehat{\theta}_i(t_i') = \theta_i\}\}.$$

Intuitively, Ξ_i is a set of payoff types that cannot be distinguished on the particular (interim) type space $\mathscr{T}$.

Fix a player i and any $t_i \in T_i$, and let

$$\Psi_i = \{\theta_i \mid \exists t_i' \in H_i^*(t_i) \text{ with } \widehat{\theta}_i(t_i') = \theta_i\}.$$

Suppose $t_i' \sim_{AM}^{\mathscr{T}} t_i$. We know that for every $\mathscr{H}^*$-measurable $\tilde{y}_i$, $\tilde{y}_i \succ_{t_i}^{\mathscr{T}} \bar{y} \Rightarrow \tilde{y}_i \succeq_{t_i'}^{\mathscr{T}} \bar{y}$. Observe that each $t_i' \in H_i^*(t_i)$ must have the same support on elements of $\mathscr{H}_{-i}^*$. Pick any t_{-i}^* such that $\widehat{\pi}_i(H_{-i}^*(t_{-i}^*) | t_i') > 0$ for all $t_i' \in H_i^*(t_i)$. Consider $\lambda_i^{\theta_i,\Psi_i}(\Psi_{-i}^{\Psi_i})$ that equals the uniform lottery everywhere except on t_{-i} with $t_j \sim_{AM}^{\mathscr{T}} t_j^*$ for all $j \neq i$, i.e.,

$$\tilde{y}_i(t_{-i}) = \bar{y} \text{ if not } t_j \sim_{AM}^{\mathscr{T}} t_j^* \text{ for some } j.$$

Note that $\tilde{y}_i$ is $\mathscr{H}^*$-measurable. Now let $\Psi_j = \{\theta_j \mid \exists t_j' \in H_j^*(t_j^*) \text{ with } \widehat{\theta}_j(t_j') = \theta_j\}$ and observe that by construction, Ψ_i is not separated by Ψ_{-i}. Thus Ξ is mutually inseparable. □

The proof of Proposition 4 follows directly from the monotone behavior of the following auxiliary sets related to the inseparable sets. In Section 2.3 we define a sequence

of inseparable sets, $\{\Xi^k\}_{k=0}^{\infty} = \{(\Xi_1^k, \ldots, \Xi_I^k)\}_{k=0}^{\infty}$, where the $k+1$st level of sets is determined by an inductive step:

$$\Xi_i^{k+1} = \{\Psi_i \in \Xi_i^k \mid \Psi_{-i} \text{ does not separate } \Psi_i, \text{ for some } \Psi_{-i} \in \Xi_{-i}^k\}. \tag{36}$$

For our monotonicity result, it is useful to simply fix a sequence of sets for all agents except i:

$$\{\Sigma_{-i}^k\}_{k=0}^{\infty} = \{(\Sigma_1^k, \ldots, \Sigma_{i-1}^k, \Sigma_{i+1}^k, \ldots, \Sigma_I^k)\}_{k=0}^{\infty}$$

such that the sequence satisfies the inclusion property

$$\Sigma_j^{k+1} \subseteq \Sigma_j^k,$$

but without necessarily coming from the separation property as Ξ_j^{k+1} in (36). However, for agent i, Σ_i^k is generated by the separation property relative to the sequence $\{\Sigma_{-i}^k\}_{k=0}^{\infty}$. In particular, $\Sigma_i^0 = 2^{\Theta_i} \setminus \varnothing$ and

$$\Sigma_i^{k+1} \triangleq \{\Psi_i \in \Sigma_i^k \mid \Psi_{-i} \text{ does not separate } \Psi_i, \text{ for some } \Psi_{-i} \in \Sigma_{-i}^k\}$$

and the resulting limit set is defined by

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

Now we consider two sequences of sets for all agents i, $\{\widehat{\Sigma}_{-i}^k\}_{k=0}^{\infty}$ and $\{\Sigma_{-i}^k\}_{k=0}^{\infty}$, such that one sequence is nested in the other, or for all k, $\widehat{\Sigma}_{-i}^k \subseteq \Sigma_{-i}^k$. We then compare the resulting limit set for agent i with respect to Σ_{-i}^k and $\widehat{\Sigma}_{-i}^k$ respectively. Correspondingly, we denote the respective limit sets of agent i by Σ_i^* and $\widehat{\Sigma}_i^*$.

Lemma 11 (Monotonicity I). *If for all k, $\widehat{\Sigma}_{-i}^k \subseteq \Sigma_{-i}^k$, then $\widehat{\Sigma}_i^* \subseteq \Sigma_i^*$.*

Proof. It suffices to show that for all k, $\widehat{\Sigma}_i^k \subseteq \Sigma_i^k$. The proof is by induction. By construction it is true for $k = 0$. Suppose now that it holds for k and we want to establish that it holds for $k+1$. By assumption, $\widehat{\Sigma}_i^k \subseteq \Sigma_i^k$ and hence consider a set $\Psi_i \in \Sigma_i^k \cap \widehat{\Sigma}_i^k$. Now suppose that $\Psi_i \in \widehat{\Sigma}_i^{k+1}$ and we want to show that $\Psi_i \in \Sigma_i^{k+1}$. We observe that if $\Psi_i \in \widehat{\Sigma}_i^{k+1}$, then there exists some $\Psi_{-i} \in \Sigma_{-i}^k$ such that Ψ_{-i} does not separate Ψ_i. But by assumption the set $\Psi_{-i} \in \Sigma_{-i}^k$, and hence it follows that $\Psi_i \in \Sigma_i^{k+1}$ as well. □

Lemma 12 (Monotonicity II). *If $\widehat{\gamma}_i > \gamma_i$, then for all k, $\Sigma_i^k \subseteq \widehat{\Sigma}_i^k$.*

Proof. The proof is by induction. By construction it is true for $k = 0$. Suppose now that it holds for k and we want to establish that it holds for $k+1$. By assumption, $\Sigma_i^k \subseteq \widehat{\Sigma}_i^k$ and hence consider a set $\Psi_i \in \Sigma_i^k \cap \widehat{\Sigma}_i^k$. Now suppose that $\Psi_i \in \Sigma_i^{k+1}$ and we want to show that $\Psi_i \in \widehat{\Sigma}_i^{k+1}$. We observe that if $\Psi_i \in \Sigma_i^{k+1}$, then there exists some $\Psi_{-i} \in \Sigma_{-i}^k$

such that Ψ_{-i} does not separate Ψ_i. In other words, there exists for every $\theta_i \in \Psi_i$ a belief $\lambda_i(\cdot \mid \theta_i) \in \Delta(\Psi_{-i})$ such that for all $x \in X$ and all $\theta_i', \theta_i'' \in \Psi_i$,

$$(1-\gamma_i)v_i(x,\theta_i') + \gamma_i \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i} \mid \theta_i') w_i(x,\theta_{-i}) = (1-\gamma_i)v_i(x,\theta_i'') + \gamma_i \sum_{\theta_{-i}\in\Theta_{-i}} \lambda_i(\theta_{-i} \mid \theta_i'') w_i(x,\theta_{-i}).$$

As the interdependent utility $w_i(\cdot)$ does not depend on θ_i, we can rewrite the equality as

$$(1-\gamma_i)(v_i(x,\theta_i') - v_i(x,\theta_i'')) = \gamma_i \sum_{\theta_{-i}\in\Theta_{-i}} (\lambda_i(\theta_{-i} \mid \theta_i'') - \lambda_i(\theta_{-i} \mid \theta_i'))w_i(x,\theta_{-i}). \tag{37}$$

Now we want to show that if $\widehat{\gamma}_i > \gamma_i$, then we can again find associated beliefs $\widehat{\lambda}_i(\theta_{-i} \mid \theta_i'')$ such that

$$(1-\widehat{\gamma}_i)(v_i(x,\theta_i') - v_i(x,\theta_i'')) = \widehat{\gamma}_i \sum_{\theta_{-i}\in\Theta_{-i}} (\widehat{\lambda}_i(\theta_{-i} \mid \theta_i'') - \widehat{\lambda}_i(\theta_{-i} \mid \theta_i'))w_i(x,\theta_{-i}). \tag{38}$$

We can easily verify that by letting for all $\theta_{-i} \in \Theta_{-i}$ the beliefs $\widehat{\lambda}_i(\theta_{-i} \mid \theta_i)$ be defined by

$$\widehat{\lambda}_i(\theta_{-i} \mid \theta_i) \triangleq \frac{(1-\widehat{\gamma}_i)\gamma_i}{\widehat{\gamma}_i(1-\gamma_i)}\lambda_i(\theta_{-i} \mid \theta_i'') + \frac{\widehat{\gamma}_i - \gamma_i}{\widehat{\gamma}_i(1-\gamma_i)}\frac{1}{(I-1)S}$$

we satisfy (38) if and only if we satisfy (37). Now since $\widehat{\gamma}_i > \gamma_i$, it follows that

$$\frac{(1-\widehat{\gamma}_i)\gamma_i}{\widehat{\gamma}_i(1-\gamma_i)} < 1,$$

and hence the conditional probability distribution $\widehat{\lambda}_i(\theta_{-i} \mid \theta_i)$ is well-defined if, as assumed, $\lambda_i(\theta_{-i} \mid \theta_i)$ is well-defined. But now it follows that $\Psi_i \in \widehat{\Sigma}_i^{k+1}$ as well. □

Proof of Proposition 4. (i) For $\gamma = \mathbf{0}$, we have by the definition of the private value utility function $v_i(\cdot)$ for all i and all θ_i and θ_i', $R_i(\theta_i, \Theta_{-i}) \cap R_i(\theta_i', \Theta_{-i}) = \varnothing$. Hence it follows that we have for all i, $\Xi_i^*(\mathbf{0}) = \{\{\theta_i^1\}, \ldots, \{\theta_i^S\}\}$. For $\gamma = \mathbf{1}$, we have by the definition of the interdependent value function $w_i(\cdot)$, for all i and all $\theta_i' \in \Theta_i$,

$$\bigcap_{\theta_i\in\Theta_i} R_i(\theta_i, \Theta_{-i}) = R_i(\theta_i', \Theta_{-i}),$$

and hence for all i, $\Xi_i^*(\mathbf{1}) = 2^{\Theta_i} \setminus \varnothing$.

(ii) It suffices to establish the result component-wise. We thus consider $\widehat{\gamma} \geq \gamma$ such that $\widehat{\gamma}_i > \gamma_i$ for some i and $\widehat{\gamma}_j = \gamma_j$ for all $j \neq i$. Now suppose that for some agent l, we have $\Xi_l^*(\gamma) \neq \Xi_l^*(\widehat{\gamma})$. Then there must be a first stage k' such that $\Xi_l^{k'}(\gamma) \neq \Xi_l^{k'}(\widehat{\gamma})$, but for all $k < k'$, we have for all l, $\Xi_l^k(\gamma) = \Xi_l^k(\widehat{\gamma})$. Now since we only changed the preferences of agent i, and k' is the first stage where the sets $\Xi_l^{k'}(\gamma)$ and $\Xi_l^{k'}(\widehat{\gamma})$ differ, it must be that $l = i$. But now it follows from Lemma 12 that $\Xi_i^{k'}(\gamma) \subset \Xi_i^{k'}(\widehat{\gamma})$. Suppose now that there is

a step $k'' > k'$ such that there exists $j \neq i$ such that $\Xi_j^{k''}(\gamma) \neq \Xi_j^{k''}(\widehat{\gamma})$, but for all $k < k''$, we have $\Xi_j^k(\gamma) = \Xi_j^k(\widehat{\gamma})$. Now we can apply Lemma 11 to conclude that $\Xi_j^k(\gamma) \subset \Xi_j^k(\widehat{\gamma})$. Now a monotonicity argument of either Lemma 11 or 12 applies at every further step along the sequence and hence we have shown that for all j, including i, we have $\Xi_j^k(\gamma) \subseteq \Xi_j^k(\widehat{\gamma})$ for all k, which establishes the result. □

References

Abreu, Dilip and Hitoshi Matsushima (1992a), "A response to Glazer and Rosenthal." *Econometrica*, 60, 1439–1442. [75]

Abreu, Dilip and Hitoshi Matsushima (1992b), "Virtual implementation in iteratively undominated strategies: Complete information." *Econometrica*, 60, 993–1008. [47, 56, 63, 65, 75]

Abreu, Dilip and Hitoshi Matsushima (1992c), "Virtual implementation in iteratively undominated strategies: Complete information." [47, 48, 51, 63, 64, 65, 68, 69, 70, 72, 73, 74, 75]

Abreu, Dilip and Hitoshi Matsushima (1994), "Exact implementation." *Journal of Economic Theory*, 64, 1–19. [74]

Artemov, Georgy, Takashi Kunimoto, and Roberto Serrano (2008), "Robust virtual implementation with incomplete information: Towards a reinterpretation of the Wilson doctrine." Discussion Paper 2007-06, Department of Economics, Brown University. [49, 71, 72]

Battigalli, Pierpaolo (1999), "Rationalizability in incomplete information games." Working Paper ECO 99/17, Department of Economics, European University Institute. [52]

Battigalli, Pierpaolo and Marciano Siniscalchi (2003), "Rationalization and incomplete information." *Advances in Theoretical Economics*, 3. [48, 52, 73]

Bergemann, Dirk and Stephen Morris (2005a), "Robust implementation: The role of large type spaces." Discussion Paper 1519, Cowles Foundation, Yale University. [75]

Bergemann, Dirk and Stephen Morris (2005b), "Robust mechanism design." *Econometrica*, 73, 1771–1813. [49, 64, 68, 76]

Bergemann, Dirk and Stephen Morris (2007), "Strategic distinguishability and robust virtual implementation." Discussion Paper 1609, Cowles Foundation, Yale University. [51, 73, 75]

Bergemann, Dirk and Stephen Morris (2008), "Robust implementation in general mechanisms." Technical Report 1666, Cowles Foundation, Yale University. [48, 52, 63]

Bergemann, Dirk and Stephen Morris (forthcoming), "Robust implementation in direct mechanisms." *Review of Economic Studies*. [49, 63, 75]

Brandenburger, Adam and Eddie Dekel (1987), "Rationalizability and correlated equilibria." *Econometrica*, 55, 1391–1402. [48, 73]

Chung, Kim-Sau and J. C. Ely (2001), "Efficient and dominance solvable auctions with interdependent valuations." Unpublished paper, Department of Economics, University of Minnesota. [74]

Chung, Kim-Sau and J. C. Ely (2007), "Foundations of dominant-strategy mechanisms." *Review of Economic Studies*, 74, 447–476. [49]

Crémer, Jacques and Richard P. McLean (1985), "Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent." *Econometrica*, 53, 345–361. [49]

Dasgupta, Partha, Peter Hammond, and Eric Maskin (1979), "The implementation of social choice rules: Some general results on incentive compatibility." *Review of Economic Studies*, 46, 185–216. [64]

Glazer, Jacob and Robert W. Rosenthal (1992), "A note on Abreu-Matsushima mechanisms." *Econometrica*, 60, 1435–1438. [75]

Harsanyi, John C. (1967–68), "Games with incomplete information played by Bayesian players, I, II, III." *Management Science*, 14, 159–182, 320–334, 486–502. [69]

Heifetz, Aviad and Zvika Neeman (2006), "On the generic (im)possibility of full surplus extraction in mechanism design." *Econometrica*, 74, 213–233. [49]

Jackson, Mathew O. (1991), "Bayesian implementation." *Econometrica*, 59, 461–477. [75]

Jackson, Matthew O. (1992), "Implementation in undominated strategies: A look at bounded mechanisms." *Review of Economic Studies*, 59, 757–775. [75]

Jehiel, Philippe, Moritz Meyer-ter Vehn, Benny Moldovanu, and William R. Zame (2006), "The limits of ex post implementation." *Econometrica*, 74, 585–610. [49]

Ledyard, John O. (1979), "Dominant strategy mechanisms and incomplete information." In *Aggregation and Revelation of Preferences* (Jean-Jacques Laffont, ed.), 309–319, North-Holland, Amsterdam. [64]

Mertens, Jean-François and Shmuel Zamir (1985), "Formulation of Bayesian analysis for games with incomplete information." *International Journal of Game Theory*, 14, 1–29. [69]

Morris, Stephen (1994), "Trade with heterogeneous prior beliefs and asymmetric information." *Econometrica*, 62, 1327–1347. [79]

Myerson, Roger B. (1981), "Optimal auction design." *Mathematics of Operations Research*, 6, 58–73. [49]

Neeman, Zvika (2004), "The relevance of private information in mechanism design." *Journal of Economic Theory*, 117, 55–77. [49]

Postlewaite, Andrew and David Schmeidler (1986), "Implementation in differential information economies." *Journal of Economic Theory*, 39, 14–33. [75]

Samet, Dov (1998), "Common priors and separation of convex sets." *Games and Economic Behavior*, 24, 172–174. [78, 79]

Sefton, Martin and Abdullah Yavas (1996), "Abreu-Matsushima mechanisms: Experimental evidence." *Games and Economic Behavior*, 16, 280–302. [75]

Serrano, Roberto and Rajiv Vohra (2001), "Some limitations of virtual Bayesian implementation." *Econometrica*, 69, 785–792. [48, 73]

Serrano, Roberto and Rajiv Vohra (2005), "A characterization of virtual Bayesian implementation." *Games and Economic Behavior*, 50, 312–331. [48, 51, 73]

Submitted 2008-4-24. Final version accepted 2009-1-16. Available online 2009-1-16.