1-2020

Rationalizable Incentives: Interim Implementation of Sets in Rationalizable Strategies

Takashi KUNIMOTO
Singapore Management University, tkunimoto@smu.edu.sg

Roberto SERRANO

SMU ECONOMICS & STATISTICS

Rationalizable Incentives: Interim Implementation of Sets in Rationalizable Strategies

Takashi Kunimoto, Roberto Serrano

January 2020

Paper No. 04-2020

Rationalizable Incentives: Interim Implementation of Sets in Rationalizable Strategies*

Takashi Kunimoto† and Roberto Serrano‡

This Version: January 2020

Abstract

This paper investigates rationalizable implementation of social choice sets (SCSs) in incomplete information environments. We identify rationalizable incentive compatibility (RIC) as its key condition, argue by means of example that RIC is strictly weaker than the standard Bayesian incentive compatibility (BIC), and show that RIC reduces to BIC when we only consider single-valued SCSs (i.e., social choice functions or SCFs). We next identify additional necessary conditions and, essentially closing the gap between necessity and sufficiency, obtain a sufficiency result for rationalizable implementation in general environments. We also characterize a well-studied class of economic environments in which RIC is essentially the only condition needed for rationalizable implementation. Considering SCFs, we show that interim rationalizable monotonicity, found in the literature, is not necessary for rationalizable implementation, as had been previously claimed.

JEL Classification: C72, D78, D82.

Keywords: rationalizable incentive compatibility, Bayesian incentive compatibility, uniform Bayesian monotonicity, interim rationalizable monotonicity, implementation, rationalizability.

1 Introduction

The theory of incentives is one of the cornerstones of modern economic theory. In it, a central condition is Bayesian incentive compatibility (BIC), viewed as a

*We owe special thanks to Rene Saran for useful comments and corrections. All remaining errors are our own.

†School of Economics, Singapore Management University, Singapore; tkunimoto@smu.edu.sg

‡Department of Economics, Brown University, Providence, RI, USA; roberto_serrano@brown.edu

minimal condition necessary for the implementation of any set of rules or contracts under incomplete information.[1] BIC stipulates that, in the direct mechanism based on a given rule, truth-telling be a best response for every type to the belief that the others are also telling the truth. Indeed, in trying to elicit the private information held by a group of agents, the mechanism designer should at least hope that, under common knowledge of rationality, if all but one of the agents are going to be truthful, so will be the remaining agent; this is the rationale for BIC as a minimal desideratum. As such, BIC is an equilibrium condition, based on the rational-expectations assumption, by which all agents share exactly the same belief about how the others will play, i.e., the truthful equilibrium belief. It turns out that the restriction imposed by BIC can sometimes be quite severe, being responsible for impossibility results in some settings.

Suppose instead that, although the mechanism designer continues to assume that rationality is commonly known by the agents, she does not insist on the rational-expectations assumption. Under incomplete information, this means that she expects the agents to use (interim) rationalizable strategies.[2] In general, if the designer's goals are summarized by a social choice set (SCS), she would seek to design a mechanism whose set of outcomes resulting from agents choosing rationalizable messages, will coincide (or at least be a subset of) the SCS of interest.[3] These would correspond, respectively, to the notions of full (weak) implementation in rationalizable strategies.

The first main result of this paper is the identification of a weakening of BIC, which we term *rationalizable incentive compatibility (RIC)*, that is necessary for full or weak implementation of SCSs in interim rationalizable strategies (Theorem 1). The definition of RIC may seem complicated at first blush, but it is very simple conceptually. It requires that truth-telling, rather than being a Bayesian equilibrium, be a rationalizable profile in the direct mechanism associated with an extended SCF whose domain is a suitably defined expanded type space (the outcomes assigned by the extended SCF over the expanded type space, when mapped back to the original type space, coincide with the outcomes prescribed by the original SCS). The proof of Theorem 1 expresses a new kind of "revelation principle" once we work with the expanded type space, a principle that is consistent with the logic of rationalizability. Namely, each expanded type consists of a type in the orig-

[1] See, e.g., Dasgupta, Hammond, and Maskin (1979), Myerson (1979, 1981), d'Aspremont and Gerard-Varet (1979), Green and Laffont (1979), and Harris and Townsend (1981) for original contributions to this fundamental idea.

[2] See Bernheim (1984), Pearce (1984), Brandenburger and Dekel (1987), and Lipman (1994) for the formalization of the idea of rationalizability; in this paper, we use the interim extension to games with incomplete information of Dekel, Fudenberg, and Morris (2007).

[3] If her goal is a unique outcome in each state, this is described by a social choice function (SCF).

inal type space and a rationalizable action in the implementing mechanism. Then, the "rationalizable mediator" recommends to each expanded type that it behave as its true original type and choose one of its rationalizable actions, and each expanded type obeys the recommendation. Notice how following the recommendation is a best response for each expanded type, given the beliefs that supported the recommended action as rationalizable in the implementing mechanism in the first place.

It is easy to see that RIC is weaker than BIC in general for SCSs, and we show by example that it is strictly weaker. On the other hand, RIC reduces to BIC for SCFs, because the beliefs to which agents best-respond can be collapsed to a singleton, given the unique outcome prescribed by the SCF. Thus, to appreciate the full power of RIC, the set-valuedness of the solution is underscored: the mechanism designer and the agents realize that, instead of making pointwise predictions – as one would do under an equilibrium logic –, they are forced to be more flexible and accept a set of outcomes that may happen, supported by rationalizable messages. Given that the rational-expectations assumption is not invoked, the designer and the agents alike are bound to be "wrong" about the actual implementation of a specific outcome, but they still accept the fact that agents choose whatever message they are choosing, as it is a best response to other agents' rationalizable messages. Notice how this is very different from the implementation of an SCS in Bayesian equilibrium, for which each SCF in the SCS can be implemented with complete independence of the rest of the set. With rationalizability and set-valuedness, implementation theory should identify the set of things that *might happen*, as opposed to the set of things that *will happen*, the latter being the view held when one assumes equilibrium theories. This subtle conceptual distinction becomes blurred if one insists on SCFs. Therefore, the weakening of the incentive constraints that entails the switch from BIC to RIC may allow for more permissive results, but they will have to be understood following this slightly different interpretation.

Our next result (Theorem 2) is not different from the analogous result for Bayesian equilibrium. It shows that if an SCS is weakly or fully implementable in rationalizable strategies, it must satisfy *closure* with respect to the concatenation of common-knowledge events. No new outcomes should be added because of any extra correlation between two such events. As in Bayesian implementation, this also applies to rationalizable strategies, which should not depend upon such correlations.

Next, the paper turns to full implementation, and identifies an additional necessary condition. We identify a new condition, *uniform Bayesian monotonicity* (UBM), which is an extension of the uniform monotonicity of Kunimoto and Serrano (2019) to incomplete-information environments. Compared to Bayesian monotonicity – BM – (Postlewaite and Schmeidler (1986), Palfrey and Srivas-

tava (1989), Jackson (1991)), UBM just changes the order of a key quantifier in the preference nestedness requirement. That is, UBM is to BM just like uniform monotonicity is to Maskin monotonicity (Maskin (1999)) in settings with complete information; importantly, UBM is weaker than BM and reduces to it for SCFs. Theorem 3 shows UBM to be necessary for full interim implementation in rationalizable strategies.

In Theorem 4, we show that the three necessary conditions (RIC, UBM, and closure), along with three additional regularity conditions often satisfied in many environments, are also sufficient for full interim rationalizable implementation. As in Kunimoto and Serrano (2019), the method of proof is a mechanism where a modulo game is centrally featured. Among the things agents should announce in the mechanism, one item is a vote for a king to be elected. The modulo game counts these votes and elects the king, and the implemented outcome is the one chosen by the king. This construction allows for the entire SCS to happen under rationalizable play, as each agent holds optimistic beliefs thinking that the king will announce the outcome that is top-ranked by the agent in the SCS.

Our general sufficiency theorem just described essentially closes the gap between necessity and sufficiency, but in doing so, it relies on an abstract mechanism that some may view as unnatural. In response, we study a wide class of economic environments, where less abstract mechanisms can be used. In particular, we show in Theorem 5 how in certain economic environments, such as an independent private-values auction, the only condition that characterizes implementation in rationalizable strategies is RIC. We illustrate our approach with the analysis of such settings in Battigalli and Siniscalchi (2003). The result can be extended to some settings with multidimensional signals, and in this context, we also discuss a general impossibility result of Jehiel and Moldovanu (2001).

Our last section is devoted to SCFs, which is the object that has been studied by previous papers dealing with iteratively undominated strategies (Abreu and Matsushima (1992)) or rationalizability (Bergemann and Morris (2008), Oury and Tercieux (2012)).[4] In particular, we discuss the connections between our conditions and interim rationalizable monotonicity (IRM), which appears in those works. In particular, we show that IRM is not necessary for rationalizable implementation. In the process, we also discuss the role of finite mechanisms, in order to understand the scope of improvement for sufficiency results.

The rest of the paper is organized as follows. In Section 2, we introduce the general notation needed for the paper. In Section 3, we propose the concept of interim implementation in rationalizable strategies. In Section 4, we illustrate the implications of rationalizable implementation via an example, which we keep revis-

[4]The Abrue-Matsushima paper uses virtual implementation, which has led to the development of an interesting literature. Here, we concentrate on exact implementation.

iting later to illustrate our concepts. In Section 5, we identify necessary conditions for implementation in rationalizable strategies. Section 6 provides the sufficiency result for rationalizable implementation in general environments. In Section 7, we study economic environments. Section 8 focuses on SCFs, and Section 9 concludes. In the appendix, we provide all the omitted proofs of results in the paper and extend our sufficiency result to environments with even weaker assumptions.

2 Preliminaries

Let $N = \{1, \ldots, n\}$ denote the finite set of agents and T_i be a finite set of types of agent i.[5] Let $T \equiv T_1 \times \cdots \times T_n$, and $T_{-i} \equiv T_1 \times \cdots \times T_{i-1} \times T_{i+1} \times \cdots \times T_n$.[6] Let $\Delta(T_{-i})$ denote the set of probability distributions over T_{-i}. Each agent i has a system of "interim" beliefs that is expressed as a function $\pi_i : T_i \to \Delta(T_{-i})$. Then, we call $(T_i, \pi_i)_{i \in N}$ a *type space*. Let A denote a finite set of pure outcomes, which are assumed to be independent of the information state. Let $\Delta(A)$ be the set of probability distributions over A. Agent i's state dependent von Neumann-Morgenstern utility function is denoted $u_i : \Delta(A) \times T \to \mathbb{R}$. We can now define an *environment* as $\mathcal{E} = (A, \{u_i, T_i, \pi_i\}_{i \in N})$.

An event $E = E_1 \times \cdots \times E_n \subseteq T$ is said to be *belief-closed* if, for each $i \in N$ and $t_i \in E_i$, we have $\pi_i(t_i)[E_{-i}] = 1$. In words, if an event E is a belief-closed subspace, it is commonly certain among the agents that E obtains. The environment is implicitly understood to be belief-closed among the agents. Suppose the planner (or mechanism designer) cares about the set of type profiles $T^* \subseteq T$. This paper takes T^* as an arbitrary belief-closed subspace of $(T_i, \pi_i)_{i \in N}$. Let $T_i^* \subseteq T_i$ comprise the set of types t_i such that there exists t_{-i} with $(t_i, t_{-i}) \in T^*$. For example, Jackson (1991) assumes that all agents have a common support prior over T. Then, T^* is interpreted as the set of profiles of types to which agents assign strictly positive probability.

A (stochastic) *social choice function* (SCF) is a single-valued function $f : T \to \Delta(A)$. Let $\mathbb{F} = \{f |\ f : T \to \Delta^*(A)\}$ be the set of SCFs, mapping into $\Delta^*(A)$, defined as a finite subset of $\Delta(A)$. The finiteness of $\mathbb{F}$ is imposed simply to avoid measurability issues in the main text.

A social choice set (SCS) F is a nonempty compact subset of $\mathbb{F}$. We say that two SCSs F and H are *equivalent* ($F \approx H$) if there exists a bijection $\xi : F \to H$

[5] In the appendix (Section A.3), we discuss the extension to a *Polish* space T_i associated with its Borel sigma-algebra $\mathscr{T}_i$. In particular, we can take T_i to be a compact subset of a Euclidean space. This is especially relevant when we discuss economic environments in Section 7. In the same section of the appendix, we also extend the analysis to infinite sets of outcomes and of social choice functions, whose definitions follow shortly.

[6] Similar notation will be used for products of other sets.

such that for every $f \in F$ and every $h \in H$ satisfying $h = \xi(f)$, $f(t) = h(t)$ for all $t \in T^*$. This means that the two SCSs "coincide" for every $t \in T^*$. For an SCF f, the interim expected utility of agent i of type t_i, who pretends to be of type t_i', is defined as:

$$U_i(f; t_i'|t_i) \equiv \sum_{t_{-i} \in T_{-i}} \pi_i(t_i)[t_{-i}] u_i(f(t_i', t_{-i})); (t_i, t_{-i})).$$

Let $U_i(f|t_i) = U_i(f; t_i|t_i)$.

A *mechanism* (or *game form*) $\Gamma = ((M_i)_{i \in N}, g)$ describes: (i) a nonempty countable message space M_i for each agent i, and (ii) an outcome function $g : M \to \Delta(A)$, where $M = \times_{i \in N} M_i$.[7] Let $\Gamma^{DR} = ((T_i)_{i \in N}, f)$ denote the *direct revelation* mechanism associated with an SCF f, i.e., a mechanism where $M_i = T_i$ for all i and $g = f$.

We close the section on preliminaries with the important notion of a deception. A *deception* is a collection $\beta = (\beta_i)_{i \in N}$, where each $\beta_i : T_i \to T_i$, and there exist $i \in N$ and $t_i \in T_i$ for whom $\beta_i(t_i) \neq t_i$. Write $\beta(t) = (\beta_1(t_1), \ldots, \beta_n(t_n))$, and for an SCF f, write $f \circ \beta$ to denote the SCF f garbled by the deception β, i.e., $f \circ \beta(t) = f(\beta(t))$ for all $t \in T$.

3 Implementation in Rationalizable Strategies

We adopt *interim correlated rationalizability* (Dekel, Fudenberg, and Morris (2007)) as a solution concept and investigate the implications of implementation in interim correlated "rationalizable" strategies. We fix a mechanism $\Gamma = (M, g)$ and define a message correspondence profile $S = (S_1, \ldots, S_n)$, where each $S_i : T_i \to 2^{M_i}$, and we write $\mathcal{S}$ for the collection of message correspondence profiles. The collection $\mathcal{S}$ is a lattice with the natural ordering of set inclusion: $S \leq S'$ if $S_i(t_i) \subseteq S_i'(t_i)$ for all $i \in N$ and $t_i \in T_i$. The largest element is $\bar{S} = (\bar{S}_1, \ldots, \bar{S}_n)$, where $\bar{S}_i(t_i) = M_i$ for each $i \in N$ and $t_i \in T_i$. The smallest element is $\underline{S} = (\underline{S}_1, \ldots, \underline{S}_n)$, where $\underline{S}_i(t_i) = \emptyset$ for each $i \in N$ and $t_i \in T_i$.

We define an operator b to iteratively eliminate never best responses. The operator $b : \mathcal{S} \to \mathcal{S}$ is thus defined as: for every $i \in N$ and $t_i \in T_i$,

$$b_i(S)[t_i] \equiv \left\{ m_i \;\middle|\; \begin{array}{l} \exists \lambda_i \in \Delta(T_{-i} \times M_{-i}) \text{ such that} \\ (1) \lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}(t_{-i}); \\ (2) \sum_{m_{-i}} \lambda_i(t_{-i}, m_{-i}) = \pi_i(t_i)[t_{-i}]; \\ (3) m_i \in \arg\max_{m_i'} \sum_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}); t_i, t_{-i}) \end{array} \right\}.$$

[7] In the appendix (Section A.3), we also discuss how the analysis can be extended to a more general class of mechanisms.

Observe that b is increasing by definition: i.e., $S \leq S' \Rightarrow b(S) \leq b(S')$. By Tarski's fixed-point theorem, there is a largest fixed point of b, which we label $S^{\Gamma(T)}$. Thus, (i) $b(S^{\Gamma(T)}) = S^{\Gamma(T)}$ and (ii) $b(S) = S \Rightarrow S \leq S^{\Gamma(T)}$. We can also construct the fixed point $S^{\Gamma(T)}$ by starting with $\bar{S}$ – the largest element of the lattice – and iteratively applying the operator b. If the message sets are finite, we have

$$S_i^{\Gamma(T)}(t_i) \equiv \bigcap_{k \geq 1} b_i \left(b^k \left(\bar{S} \right) \right) [t_i]$$

for each $i \in N$ and $t_i \in T_i$. However, since the mechanism Γ may be infinite, transfinite induction may be necessary to reach the fixed point. It is useful to define

$$S_i^{\Gamma(T),k}(t_i) \equiv b_i \left(b^{k-1} \left(\bar{S} \right) \right) [t_i],$$

using transfinite induction if necessary. Thus, $S_i^{\Gamma(T)}(t_i)$ are the sets of messages surviving (transfinite) iterated deletion of never best responses of type t_i of agent i. We denote by σ_i a selection from $S_i^{\Gamma(T)}$ and call it a rationalizable strategy of agent i. We recall the following structure of $S^{\Gamma(T)}$:

$$S^{\Gamma(T)} = \prod_{i \in N} S_i^{\Gamma(T)}.$$

Next, we provide the definitions of *rationalizable implementation* that we use in the paper.

Definition 1 (Full Implementation in Rationalizable Strategies) *An SCS F is **fully implementable in rationalizable strategies** if there exists a mechanism $\Gamma = (M, g)$ with the following three properties: (i) for each $t \in T^*$,*

$$\bigcup_{m \in S^{\Gamma(T)}(t)} \{g(m)\} = F(t),$$

(ii) for any $i \in N$, $t_i \in T_i$, and $\sigma_i \in S_i^{\Gamma(T)}$, there exist a belief $\lambda_i^{\sigma_i(t_i)} \in \Delta(T_{-i} \times M_{-i})$ and profile of pure strategies $\sigma_{-i} \in S_{-i}^{\Gamma(T)}$, such that $\lambda_i^{\sigma_i(t_i)}(t_{-i}, \sigma_{-i}(t_{-i})) = 1$ for each $t_{-i} \in T_{-i}$, and $\sigma_i(t_i)$ is a best response against $\lambda_i^{\sigma_i(t_i)}$, and (iii) for any $\sigma', \sigma'' \in S^{\Gamma(T)}$ and $T', T'' \subseteq T$ for which $T' \cup T'' = T$ and $T' \cap T'' = \emptyset$, there exists $f \in F$ such that for any $t \in T^*$,

$$f(t) = \begin{cases} g(\sigma'(t)) & \text{if } t \in T' \\ g(\sigma''(t)) & \text{if } t \in T'' \end{cases}$$

Condition (i) is the usual requirement for full implementation. That is, over the states the designer cares about, each outcome that corresponds to rationalizable message profiles must be in the SCS of interest, and vice versa, for each SCF in the

SCS of interest, there exist rationalizable messages that yield it as their outcome. Condition (ii) is a slight strengthening, and it requires that every rationalizable strategy can be made a best response to some "degenerate" belief regarding other players' strategies. We note that this second requirement is inconsequential when we only consider SCFs. Condition (iii) is a weak regularity requirement, which imposes no restriction on rationalizably implementable SCSs as long as we care about equivalent SCSs.

A definition that is weaker than full implementation relaxes Condition (i) to require only its former part:

Definition 2 (Weak Implementation in Rationalizable Strategies) *An SCS F is **weakly implementable in rationalizable strategies** if there exists a mechanism $\Gamma = (M, g)$ with the following three properties: (i) for each $t \in T^*$,*

$$\emptyset \neq \bigcup_{m \in S^{\Gamma(T)}(t)} \{g(m)\} \subseteq F(t),$$

and conditions (ii) and (iii) as above.

It is clear that, if one is interested in implementing an SCF, both notions of implementation just presented coincide.

4 An Example

In order to begin illustrating the different concepts in the paper, we consider the following example, an incomplete-information variant of the main example in Kunimoto and Serrano (2019). There are two agents $N = \{1, 2\}$; two states $\{\alpha, \beta\}$ and a finite number of pure outcomes $A = \{a_1, a_2, \ldots, a_K\}$ where $K \geq 4$. For simplicity, we assume that the set of lotteries over A includes each pure outcome as a degenerate lottery and it is a finite set. We denote by $\Delta^F(A)$ such a set of lotteries over A. Assume that agent 1 is uninformed of the state and agent 2 is informed of the state. Accordingly, we define $T_2 = \{t_\alpha, t_\beta\}$ as the set of types for agent 2 such that agent 2 of type t_α knows that the state is α and type t_β knows that the state is β. Assume also that agent 1 believes with probability q_α that the state is α and with probability $(1 - q_\alpha)$ that it is β where $q_\alpha \in (0, 1)$.

Agent 1's utility function has the following features: (1) $u_1(a_k) = u_1(a_k; \alpha) = u_1(a_k; \beta)$ for each $a_k \in A$ (state-independence) and (2)

$$u_1(a_K) > u_1(a_1) > u_1(a_2) > \cdots > u_1(a_{K-1}).$$

Agent 2's utility function in state α has the following features:

$$u_2(a_K; \alpha) > u_2(a_2; \alpha) > u_2(a_{K-1}; \alpha) > \cdots > u_2(a_1; \alpha).$$

Agent 2's utility function in state β has the following features:

$$u_2(a_K;\beta) > u_2(a_{K-1};\beta) > \cdots > u_2(a_1;\beta) > u_2(a_2;\beta).$$

We further assume that $u_2(a_k) = u_2(a_k,\alpha) = u_2(a_k;\beta)$ for any $a_k \in A\backslash\{a_2\}$ (state-independence, except for a_2).

We next discuss the value of q_α needed for the argument that follows. Define

$$\varepsilon = \min_{k\neq K} \min_{\phi\in\Delta^F(A):u_1(\phi)-u_1(a_k)>0} u_1(\phi) - u_1(a_k).$$

We choose q_α large enough so that $q_\alpha > 1/(1+\varepsilon)$.

We consider the following SCS $F = \{f_{K-1,K}, f_{K,K}\} \cup \{f_{k,K-1}\}_{k=1}^{K-1}$, where $f_{i,j}$ denotes the SCF that assigns alternative a_i in state α and a_j in state β.

We now show that the SCS F is implementable in rationalizable strategies using a finite mechanism. Consider the following mechanism $\Gamma = (M, g)$ where $M_i = \{m_i^1, m_i^2, \ldots, m_i^K\}$ for each $i = 1, 2$ and the deterministic outcome function $g(\cdot)$ is given in the table below:

$g(m)$		Agent 2						
		m_2^1	m_2^2	m_2^3	m_2^4	$\cdots$	m_2^{K-1}	m_2^K
Agent 1	m_1^1	a_1	a_1	a_{K-2}	a_{K-3}	$\cdots$	a_2	a_{K-1}
	m_1^2	a_2	a_1	a_1	a_{K-2}	$\cdots$	a_3	a_{K-1}
	m_1^3	a_3	a_2	a_1	a_1	$\cdots$	a_4	a_{K-1}
	m_1^4	a_4	a_3	a_2	a_1	$\cdots$	a_5	a_{K-1}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	m_1^{K-1}	a_1	a_{K-2}	a_{K-3}	a_{K-4}	$\cdots$	a_1	a_{K-1}
	m_1^K	a_{K-1}	a_{K-1}	a_{K-1}	a_{K-1}	$\cdots$	a_{K-1}	a_K

Fact 1 *The SCS F is fully implementable in rationalizable strategies by the finite mechanism Γ.*

Proof: Recall that we choose q_α sufficiently close to one. This means that agent 1 believes with sufficiently high probability that the state is α. Then, no message of agent 1 in the mechanism Γ is eliminated. We elaborate on this argument further: for each $k \in \{1, \ldots, K-1\}$, m_1^k is a best response to the belief that agent 2 chooses m_2^k for sure. In addition, m_1^K is a best response to the belief that agent 2 chooses m_2^K for sure.

Recall that agent 2 is informed of the state. In state α, no message of agent 2 in the mechanism $\Gamma = (M, g)$ is eliminated. Specifically, we argue as follows:

for each $k \in \{1, \ldots, K-2\}$, m_2^k can be a best response to the belief that agent 1 chooses m_1^{k+1} for sure. In addition, m_2^{K-1} can be a best response to the belief that agent 1 chooses m_1^1 for sure; and m_2^K is a best response to the belief that agent 1 chooses m_1^K for sure.

Consider now state β. In that state, message m_2^K strictly dominates all other messages, $m_2^1, \ldots, m_2^{K-1}$.

Furthermore, since q_α is high enough, agent 1 is still unable to eliminate any message, even after agent 2's first round of elimination (all messages but one in state β). It follows that no further eliminations can occur.

The preceding arguments show that $S_1^{\Gamma(T)} = M_1$, $S_{2,t_\alpha}^{\Gamma(T)} = M_2$, and $S_{2,t_\beta}^{\Gamma(T)} = \{m_2^K\}$. Therefore, all the SCFs in the SCS proposed are implementable in interim rationalizable strategies. Specifically, when agent 1 chooses m_1^K, both outcomes a_{K-1} and a_K can occur in state α, but only a_K in state β. And when agent 1 chooses any other message, any outcome in $\{a_1, a_2, \ldots, a_{K-1}\}$ can happen in state α, but only a_{K-1} in state β. We also confirm that the second requirement of implementability is satisfied. This completes the proof. ■

This example will become important for many of our conditions, and we shall revisit it multiple times in due course as a result.

5 Necessary Conditions for SCSs

5.1 Necessary Conditions for Weak or Full Implementation

To begin with, consider the traditional *incentive compatibility* condition:

Definition 3 *An SCS F satisfies* ***Bayesian incentive compatibility (BIC)*** *if, for each $f \in F$, $i \in N$, and $t_i, t_i' \in T_i^*$,*

$$U_i(f|t_i) \geq U_i(f; t_i'|t_i).$$

In addition, F satisfies ***strict-if-responsive Bayesian incentive compatibility (SIRBIC)*** *if, for every $f \in F$, the above inequality holds as a strict inequality whenever there exists $\hat{t}_{-i} \in T_{-i}$ such that $f(t_i, \hat{t}_{-i}) \neq f(t_i', \hat{t}_{-i})$; and F satisfies* ***strict*** *Bayesian incentive compatibility if, for every $f \in F$, the above inequality holds as a strict inequality whenever $t_i \neq t_i'$.*

Bayesian incentive compatibility is known to be a necessary condition for the (weak or full) implementation in Bayesian equilibrium of SCSs. The reader is referred to Jackson (1991) for this, for example. Considering again the example in Section 4, BIC is not necessary for (weak or full) implementation in rationalizable strategies:

Fact 2 *The SCS F in the example of section 4 violates Bayesian incentive compatibility.*

Proof: This is easy to see. The direct mechanism associated with any SCF is a one-agent game form; given the preferences of the informed agent, any nonconstant SCF in the SCS F violates Bayesian incentive compatibility. ■

Instead, a new condition that describes "rationalizable incentives" is called for, and that condition will become central in our paper. Indeed, we propose a weaker condition, and preparing for it, we provide the next definitions.

The tuple $\mathcal{T} = (\mathcal{T}_i, T_i, \hat{t}_i, \hat{\pi}_i)_{i\in N}$ is said to be an *expanded type space* on $(T_i, \pi_i)_{i\in N}$ if, for each $i \in N$, $\hat{t}_i : \mathcal{T}_i \to T_i$ is an onto mapping and for each $\tau_i \in \mathcal{T}_i$ and $t_{-i} \in T_{-i}$, $\hat{\pi}_i : \mathcal{T}_i \to \Delta(\mathcal{T}_{-i})$ satisfies that

$$\sum_{\tau_{-i}:\hat{t}_{-i}(\tau_{-i})=t_{-i}} \hat{\pi}_i(\hat{t}_i(\tau_i))[\tau_{-i}] = \pi_i(\hat{t}_i(\tau_i))[t_{-i}].$$

Definition 4 *An SCS F satisfies* ***rationalizable incentive compatibility (RIC)*** *if there exist an expanded type space $\mathcal{T} = (\mathcal{T}_i, T_i, \hat{t}_i, \hat{\pi}_i)_{i\in N}$ on $(T_i, \pi_i)_{i\in N}$ and an extended SCF $\tilde{f} : \mathcal{T} \to \Delta(A)$ such that the identity mapping $I : \mathcal{T} \to \mathcal{T}$ constitutes a rationalizable message profile in the associated direct mechanism $\Gamma^{DR}(\mathcal{T}, \tilde{f})$ and $\tilde{f} \circ \hat{t}^{-1} \approx F' \subseteq F$, where $\tilde{f} \circ \hat{t}^{-1}$ denotes the set of SCFs such that $\bigcup_{\tau:\hat{t}(\tau)=t}\{\tilde{f}(\tau)\}$ for each occurrence of t.*

An SCS F satisfies ***fully rationalizable incentive compatibility*** *if it satisfies RIC and $F' = F$.*

For SCFs, the RIC condition can be written as follows:

Definition 5 *An SCF f satisfies* ***rationalizable incentive compatibility (RIC)*** *if there exist an expanded type space $\mathcal{T} = (\mathcal{T}_i, T_i, \hat{t}_i, \hat{\pi}_i)_{i\in N}$ on $(T_i, \pi_i)_{i\in N}$ and an extended SCF $\tilde{f} : \mathcal{T} \to \Delta(A)$ such that the identity mapping $I : \mathcal{T} \to \mathcal{T}$ constitutes a rationalizable message profile in the associated direct mechanism $\Gamma^{DR}(\mathcal{T}, \tilde{f})$ and $\tilde{f} \circ \hat{t}^{-1} \approx f$, where $\tilde{f} \circ \hat{t}^{-1}$ denotes the set of SCFs such that $\bigcup_{\tau:\hat{t}(\tau)=t}\{\tilde{f}(\tau)\}$ for each occurrence of t.*

We establish the following observation:

Proposition 1 *An SCF f satisfies rationalizable incentive compatibility if and only if it satisfies Bayesian incentive compatibility.*

Proof: The "if" part is straightforward to show. So, we only focus on the "only if" part. To show this, suppose that the SCF f satisfies RIC, which guarantees the existence of the expanded type space $\mathcal{T} = (\mathcal{T}_i, T_i, \hat{t}_i, \hat{\pi}_i)_{i\in N}$ on $(T_i, \pi_i)_{i\in N}$, the

extended SCF $\tilde{f} : \mathcal{T} \to \Delta(A)$ such that $\tilde{f} \circ \hat{t}^{-1} \approx f$, and the fact that truth-telling is rationalizable in the associated direct mechanism $\Gamma^{DR}(\mathcal{T}, \tilde{f})$. We set $\mathcal{T}_i = T_i \times M_i$ such that M_i is an arbitrary countable set and $\hat{t}_i(t_i, m_i) = t_i$ for each $(t_i, m_i) \in T_i \times M_i$. Thus, we get that, for every agent $i \in N$, extended type $\tau_i = (t_i, m_i)$ is best-responding by telling the truth to her belief λ_i with support of the following kind: $\lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow$ for each $j \neq i$, extended type (t_j, m_j) truthfully announces (t_j, m_j). Taking into account the definition of $\hat{\pi}_i(\hat{t}_i(\tau_i))[\tau_{-i}]$, and since the outcome of truth-telling in the direct mechanism $\Gamma^{DR}(\mathcal{T}, \tilde{f})$ is $f(t)$ for every underlying type profile t, one can take, without loss of generality, a degenerate belief λ_i held by type t_i in the direct mechanism for f concentrated on the true type reports $(t_j)_{j \neq i}$, to which truth-telling is a best response. Since agent i and type t_i were chosen arbitrarily, this shows that truth-telling is a Bayesian (Nash) equilibrium in the direct mechanism for f, and hence, f satisfies BIC. ■

We next present our first main result:

Theorem 1 *If an SCS F is (weakly or fully) implementable in rationalizable strategies, there exists an equivalent SCS $\hat{F} \approx F$ such that the SCS $\hat{F}$ satisfies rationalizable incentive compatibility.*

Proof: Suppose that the SCS F is (weakly or fully) implementable in rationalizable strategies by the mechanism $\Gamma = (M, g)$. Then, for any $t \in T^*$,

$$\bigcup_{m \in S^{\Gamma(T)}(t)} \{g(m)\} = F'(t) \subseteq F(t).$$

When we require full (rather than weak) implementation, we have $F'(t) = F(t)$. For each $i \in N$, we define

$$\mathcal{T}_i = \bigcup_{t_i \in T_i} \bigcup_{m_i \in S_i^{\Gamma(T)}(t_i)} \{(t_i, m_i)\}$$

and for each $\tau_i \in \mathcal{T}_i$, there exists $t_i \in T_i$ such that $m_i \in S_i^{\Gamma(T)}(t_i)$, where $\tau_i = (t_i, m_i)$. Then, we define $\hat{t}_i(\tau_i) = t_i$ accordingly. Here $\hat{t}_i$ maps for each new augmented type the original type behind that message. This is an onto map.

Thus, constructing the beliefs $\hat{\pi}_i$ making use of the beliefs $\lambda_i(m_{-i}, t_{-i})$ held by t_i to which each message $m_i \in S^{\Gamma(T)}(t_i)$ is a best reply for type t_i, and obeying the obvious adding-up constraints from the π_i's, we have an expanded type space $\mathcal{T} = (\mathcal{T}_i, T_i, \hat{t}_i, \hat{\pi}_i)_{i \in N}$.

For each $\sigma \in S^{\Gamma(T)}$, by implementability, the outcome function of the mechanism yields an SCF $f \in F$ such that $f = g \circ \sigma$. Then, one can define an extended SCF $\tilde{f}$: for each $\tau \in \mathcal{T}$, there exists a message profile $m \in S^{\Gamma(T)}(\hat{t}(\tau))$ such that

$\tilde{f}(\tau) = g(m)$. By construction of the expanded type space $\mathcal{T}$ and the extended SCF $\tilde{f}$, we can easily see that the identity mapping $I : \mathcal{T} \to \mathcal{T}$ constitutes a rationalizable strategy profile in the associated direct mechanism $\Gamma^{DR}(\mathcal{T}, \tilde{f})$, i.e., $I \in S^{\Gamma^{DR}(\mathcal{T},\tilde{f})}$. Furthermore, defining an SCS $\hat{F}$ such that $\hat{F}(t) = \bigcup_{\tau:\hat{t}(\tau)=t}\{\tilde{f}(\tau)\}$ for any $t \in T$, by construction, we have that $\hat{F} \approx F^{'} \subseteq F$. This completes the proof. ■

Remark: Rationalizable incentive compatibility of an SCS F is equivalent to truthful implementation of $\tilde{f}$ in rationalizable strategies. That is, it requires that truth-telling be a rationalizable profile in a suitably defined direct mechanism for an extended type space. Then, note how the proof expresses a new kind of "revelation principle" once we work with the extended type space, a principle that is consistent with the logic of rationalizability. In this light, each extended type consists of a type in the original type space and a rationalizable action in the implementing mechanism. Then, the "rationalizable mediator" recommends to each extended type that it behave as its true original type and choose one of its rationalizable actions, and each extended type obeys the recommendation. Notice how following the recommendation is a best response for each extended type, given the beliefs that supported the recommended action as rationalizable in the implementing mechanism.

For instance, in the example of Section 4, we have:

Fact 3 *The SCS F in the example of Section 4 satisfies rationalizable incentive compatibility.*

This new fact follows from Theorem 1. In particular, the set of extended types for agent 1 in the example is $\{m_1^1, m_1^2, \ldots, m_1^K\}$, i.e., the set of the K actions, all of them being rationalizable. In contrast, for agent 2, the set of extended types is the following $K+1$-element set: $\{(t_\alpha, m_2^1), (t_\alpha, m_2^2), \ldots, (t_\alpha, m_2^K), (t_\beta, m_2^K)\}$.

Remark: RIC can sometimes be violated. First, by Proposition 1, if we are confined to SCFs, any SCF that violates BIC also violates RIC. For SCSs, the following is an example that violates RIC. There are two states t and $t^{'}$. Agent 1 is informed about the state, and agent 2 is not. There are three alternatives, $A = \{a, b, c\}$. The utility values for agent 1 over A are $u_1(\cdot, t) = (1, 2, 0)$ in state t and $u_1(\cdot, t^{'}) = (2, 1, 0)$ in state $t^{'}$. The utility values for agent 2 over A are all 0. Consider an SCS consisting of only two SCFs, both of which violate BIC: $F = \{f, f^{'}\}$, where $f(t) = a, f(t^{'}) = b$, $f^{'}(t) = a, f^{'}(t^{'}) = c$. We claim that there does not exist a mechanism that fully implements F in rationalizable strategies. To see this, note that agent 2 will never be able to eliminate any message, given

that he is indifferent among all outcomes. Suppose he is the "column player" in the proposed mechanism. For agent 1, the "row player," we must have that: (i) in state t, only outcome a must happen, so she must be able to rationalize an action, say the first row, where the only outcome is a; and (ii) in state $t^{'}$, she must rationalize other actions, say second row, where only b and c must feature, but then, given her preferences in state t, she could not eliminate this row to leave only the first one, because at least for a column, outcome b must show up there, which is better than a in state t.

The next necessary condition is already featured in the literature on implementation in Bayesian equilibrium (Jackson (1991)). We need some preliminaries to introduce it. For a belief-closed subspace $E \subseteq T$ and an SCS F, define

$$F(E) \equiv \left\{\alpha \in \Delta(A) \middle| \ \exists\ f \in F,\ \exists t \in E \text{ s.t. } f(t) = \alpha\right\}.$$

The reader is referred to Section 2 for the definition of belief-closedness. For two belief-closed subspaces $E, E^{'} \subseteq T$ and an SCS F, define

$$F(E \times E^{'}) \equiv \left\{(\alpha, \alpha^{'}) \in \Delta(A) \times \Delta(A) \middle| \ \exists f, f^{'} \in F,\ \exists t \in E, \exists t^{'} \in E^{'} \text{ s.t. } f(t) = \alpha \text{ and } f^{'}(t^{'}) = \alpha^{'}\right\}.$$

Definition 6 *An SCS F satisfies **closure** if, for any pair of belief-closed subspaces $E, E^{'} \subseteq T$, we have*

$$F(E \times E^{'}) = F(E) \times F(E^{'}).$$

In words, closure says that new outcomes should not be added because of any extra correlation between two belief-closed subspaces. This should apply to rationalizable strategies, which should not depend upon such correlations, and indeed, the next result shows that closure is a necessary condition for rationalizable implementation:

Theorem 2 *If an SCS F is (weakly or fully) implementable in rationalizable strategies, it satisfies closure.*

Remark: Note that every SCF trivially satisfies closure.

Proof: Suppose that the SCS F is (weakly or fully) implementable in rationalizable strategies by the mechanism $\Gamma = (M, g)$. Let E^1, E^2 be a pair of belief-closed subspaces in T. By definition, it is easy to see that

$$F\left(E^1 \times E^2\right) \subseteq F\left(E^1\right) \times F\left(E^2\right).$$

Therefore, it only remains to show the converse. That is,

$$F\left(E^1 \times E^2\right) \supseteq F\left(E^1\right) \times F\left(E^2\right).$$

By the definition of implementation, for each $i \in N$, there exists a pair of strategy profiles $\sigma_i^1 \in \Sigma_i$ and $\sigma_i^2 \in \Sigma_i$ such that $\sigma^1(t) \in S^{\Gamma(T)}(t)$ and $g(\sigma^1(t)) \in F(E^1)$ for every $t \in E^1$; and $\sigma^2(t) \in S^{\Gamma(T)}(t)$ and $g(\sigma^2(t)) \in F(E^2)$ for each $t \in E^2$. For each $i \in N$, define σ_i as follows: for each $t \in E^1 \cup E^2$,

$$\sigma_i(t) = \begin{cases} \sigma_i^1(t) & \text{if } t \in E^1 \\ \sigma_i^2(t) & \text{if } t \in E^2 \end{cases}$$

We claim that $\sigma(t) \in S^{\Gamma(T)}(t)$ for each $t \in E^1 \cup E^2$. Fix $i \in N$ and $t_i \in E_i^1 \cup E_i^2$. If $t_i \in E_i^1$, there exists $\lambda_i^1 \in \Delta(T_{-i} \times M_{-i})$ such that (i) $\lambda_i^1(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$, (ii) $\sum_{m_{-i}} \lambda_i^1(t_{-i}, m_{-i}) = \pi_i(t_i)[t_{-i}]$, and (iii)

$$\sigma_i^1(t_i) \in \arg\max_{m_i'} \sum_{t_{-i}, m_{-i}} \lambda_i^1(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}); (t_i, t_{-i}))$$

Similarly, if $t_i \in E_i^2$, there exists $\lambda_i^2 \in \Delta(T_{-i} \times M_{-i})$ such that (i) $\lambda_i^2(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$, (ii) $\sum_{m_{-i}} \lambda_i^2(t_{-i}, m_{-i}) = \pi_i(t_i)[t_{-i}]$ and (iii)

$$\sigma_i^2(t_i) \in \arg\max_{m_i'} \sum_{t_{-i}, m_{-i}} \lambda_i^2(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}); (t_i, t_{-i}))$$

Due to the construction of σ_i and the hypothesis that E^1 and E^2 each are belief-closed subspaces, there exists $\lambda_i \in \Delta(T_{-i} \times M_{-i})$ such that

$$\begin{aligned} & \sum_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i(g(\sigma_i(t_i), m_{-i}); t_i, t_{-i}) \\ = & \sum_{t_{-i}, m_{-i}} \lambda_i^1(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}); (t_i, t_{-i})) + \sum_{t_{-i}, m_{-i}} \lambda_i^2(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}); (t_i, t_{-i})). \end{aligned}$$

This implies that (i) $\lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$, (ii) $\sum_{m_{-i}} \lambda_i(t_{-i}, m_{-i}) = \pi_i(t_i)[t_{-i}]$, and (iii)

$$\sigma_i(t_i) \in \arg\max_{m_i'} \sum_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}); (t_i, t_{-i})).$$

This implies that $\sigma_i(t_i) \in S_i^{\Gamma(T)}(t_i)$ for each $t_i \in E_i^1 \cup E_i^2$. Thus, $\sigma(t) \in S^{\Gamma(T)}(t)$ for each $t \in E^1 \cup E^2$. This allows us to conclude that $g \circ (\sigma^1, \sigma^2) \in F(E^1 \cup E^2)$, which completes the proof. ■

For instance, the SCS F we consider in the example of Section 4 satisfies closure. Moreover, we have:

Fact 4 *Any SCS satisfies closure in the example of Section 4.*

This follows simply because there are no two disjoint belief-closed subspaces in the type space of the example.

5.2 An Additional Necessary Condition for Full Implementation

For full implementation, additional monotonicity conditions are generally required beyond incentive compatibility. We first recall *Bayesian monotonicity* (see Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989), and Jackson (1991)). Bayesian monotonicity is known to be a necessary condition for the full implementation of SCSs in Bayesian equilibrium.

Definition 7 *An SCS F satisfies* ***Bayesian monotonicity (BM)*** *if, for every $f \in F$ and deception β, whenever for $y : T_{-i} \to \Delta(A)$ it holds that:*

$$(*) \quad \left[\begin{array}{l} U_i(f|\tilde{t}_i) \geq U_i(y|\tilde{t}_i) \ \forall \tilde{t}_i \in T_i \\ \Rightarrow U_i(f \circ \beta|t_i) \geq U_i(y \circ \beta|t_i) \end{array} \right], \forall i \in N, \ \forall t_i \in T_i^*,$$

then $f \circ \beta \in F$.

Below, we show that BM is not necessary for full implementation in rationalizable strategies of SCSs. We present next a variant of BM, which does turn out to be necessary:

Definition 8 *An SCS F satisfies* ***uniform Bayesian monotonicity (UBM)*** *if, for every deception β, whenever for $y : T_{-i} \to \Delta(A)$ it holds that:*

$$(**) \left[\begin{array}{l} U_i(f|\tilde{t}_i) \geq U_i(y|\tilde{t}_i) \ \forall \tilde{t}_i \in T_i \\ \Rightarrow U_i(f \circ \beta|t_i) \geq U_i(y \circ \beta|t_i) \end{array} \right], \forall f \in F, \ \forall i \in N, \ \forall t_i \in T_i^*,$$

then $F \circ \beta \subseteq F$.

Remark: Note how, with respect to BM, the only difference is the order of quantifiers in the clause "$\forall f \in F$" regarding the appropriate inclusion of lower contour sets; see Kunimoto and Serrano (2019) for a similar difference between Maskin monotonicity and uniform monotonicity in complete information environments. It follows that both BM and UBM coincide if one is dealing with SCFs.

Theorem 3 *If an SCS F is fully implementable in rationalizable strategies, it satisfies uniform Bayesian monotonicity.*

Proof: Suppose that the SCS F is fully implementable in rationalizable strategies by a mechanism $\Gamma = (M, g)$. Fix a deception β satisfying condition $(**)$. We show that $F \circ \beta \subseteq F$.

We first show that $S^{\Gamma(T)} \supseteq S^{\Gamma(\beta(T))}$, the latter being the set of relabeled strategy profiles $\sigma \circ \beta$ for each rationalizable profile σ, when type t_i pretends to be type

$\beta_i(t_i)$, for all $t_i \in T_i$ and all $i \in N$. Recall that $b(S^{\Gamma(T)}) = S^{\Gamma(T)}$, i.e., $S^{\Gamma(T)}$ has the best response property. Fix $\sigma \in S^{\Gamma(T)}$ arbitrarily. This implies that for each agent $i \in N$ and type $t_i \in T_i$, there exists $\lambda_i^{\sigma_i(t_i)} \in \Delta(T_{-i} \times M_{-i})$ such that (1) $\lambda_i^{\sigma_i,t_i}(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$, (2) $\sum_{m_{-i}} \lambda_i^{\sigma_i(t_i)}(m_{-i}, t_{-i}) = \pi_i(t_i)[t_{-i}]$, and (3) for all $m_i^{'} \in M_i$,

$$\begin{aligned} & \sum_{t_{-i}, m_{-i}} \lambda_i^{\sigma_i(t_i)}(t_{-i}, m_{-i}) u_i(g(\sigma_i(t_i), m_{-i}); t_i, t_{-i}) \\ \geq & \sum_{t_{-i}, m_{-i}} \lambda_i^{\sigma_i(t_i)}(t_{-i}, m_{-i}) u_i(g(m_i^{'}, m_{-i}); t_i, t_{-i}). \end{aligned}$$

We want to show that $\sigma_i(\beta_i(t_i))$ is also a best response against the suitably relabeled $\lambda_i^{\sigma_i(\beta_i(t_i))}$. We focus on the best response property of σ_i summarized by inequality (3). By the second requirement of implementability, we can assume that there exists $\sigma_{-i} \in S_{-i}^{\Gamma(T)}$ such that $\lambda_i^{\sigma_i(t_i)}(t_{-i}, \sigma_{-i}(t_{-i})) = 1$ for each $t_{-i} \in T_{-i}$.[8] Due to the implementability of F, we have that $g \circ (\sigma_i, \sigma_{-i}) \in F$ as an SCF, which we denote by f. Define $y = g \circ (m_i^{'}, \sigma_{-i})$. Note that y is independent of agent i's type. Thus, we have

$$U_i(f|\tilde{t}_i) \geq U_i(y|\tilde{t}_i)$$

for all $\tilde{t}_i \in T_i$. Thus, using condition $(**)$, we also have

$$U_i(f \circ \beta|t_i) \geq U_i(y \circ \beta|t_i).$$

For each $(t_{-i}, m_{-i}) \in T_{-i} \times M_{-i}$, define

$$\lambda_i^{\sigma_i(\beta_i(t_i))}(t_{-i}, m_{-i}) = \begin{cases} \lambda_i^{\sigma_i(t_i)}(t_{-i}, \sigma_{-i}(\beta_{-i}(t_{-i}))) & \text{if } m_{-i} = \sigma_{-i}(\beta_{-i}(t_{-i})) \\ 0 & \text{otherwise.} \end{cases}$$

By construction, we clearly satisfy $\sum_{m_{-i}} \lambda_i^{\sigma_i(\beta_i(t_i))}(t_{-i}, m_{-i}) = \pi_i(t_i)[t_{-i}]$ for each $t_{-i} \in T_{-i}$. Then $U_i(f \circ \beta|t_i) \geq U_i(y \circ \beta|t_i)$ implies that for any $m_i^{'} \in M_i$,

$$\begin{aligned} & \sum_{t_{-i}, m_{-i}} \lambda_i^{\sigma_i(\beta_i(t_i))}(t_{-i}, m_{-i}) u_i(g(\sigma_i(\beta_i(t_i)), m_{-i}); t_i, t_{-i}) \\ \geq & \sum_{t_{-i}, m_{-i}} \lambda_i^{\sigma_i(\beta_i(t_i))}(t_{-i}, m_{-i}) u_i(g(m_i^{'}, m_{-i}); t_i, t_{-i}). \end{aligned}$$

This shows that $\sigma_i \circ \beta_i$ is a best response against the suitably relabeled belief $\lambda_i^{\sigma_i(\beta_i(t_i))}$. Since the choice of agent i and $\sigma_i \in S_i^{\Gamma(T)}$ is arbitrary, we can conclude that $S^{\Gamma(\beta(T))} \subseteq S^{\Gamma(T)}$. Finally, by full implementability, this implies that

$$F(t) = \bigcup_{m \in S^{\Gamma(T)}(t)} \{g(m)\} \supseteq \bigcup_{m \in S^{\Gamma(\beta(T))}(t)} \{g(m)\} = F(\beta(t)).$$

[8] This requirement is unnecessary for the case of SCFs.

The proof is thus complete. ■

We can now revisit again the example we discussed in Section 4, in order to illustrate the concepts just introduced. Here is a new result for it:

Fact 5 *The SCS F in the example of Section 4 violates Bayesian monotonicity.*

Proof: Consider a deception $\gamma : T_2 \to T_2$ such that $\gamma(t_\alpha) = \gamma(t_\beta) = t_\alpha$. Then, we have that for $k \neq 2, K-1, K$, $f_{k,K-1}(\tilde{t}) = a_k$ for each $\tilde{t} \in T_2$ so that $f_{k,K-1} \circ \gamma \notin F$. We want to show that condition $(*)$ is satisfied for $f_{k,K-1}$ and γ, contradicting the conclusion of Bayesian monotonicity.

We check the following:

$$U_2(f_{k,K-1}|t_\alpha) = u_2(a_k, \alpha) \text{ and } U_2(f_{k,K-1}|t_\beta) = u_2(a_{K-1}, \beta).$$

The requirement that $U_2(f_{k,K-1}|\tilde{t}) \geq U_2(y|\tilde{t})$ for all $\tilde{t} \in T_2 = \{t_\alpha, t_\beta\}$ concludes that such y's must be lotteries that assign positive probability only to $\{a_1, a_3, a_4, \ldots, a_{k-1}, a_k\}$. (Indeed, the differences in utility values for agent 2 in state α can be chosen arbitrarily large for this to be true.) We thus get

$$\begin{aligned} U_2(f_{k,K-1} \circ \gamma|t_\alpha) &= u_2(a_k, \alpha) \geq U_2(y \circ \gamma|t_\alpha) \\ U_2(f_{k,K-1} \circ \gamma|t_\beta) &= u_2(a_k, \beta) \geq U_2(y \circ \gamma|t_\beta), \end{aligned}$$

This shows that condition $(*)$ holds for agent 2.

We now move on to agent 1. We compute the following:

$$\begin{aligned} U_1(f_{k,K-1}) &= q_\alpha u_1(a_k) + (1-q_\alpha)u_1(a_{K-1}) \\ U_1(y) &= q_\alpha u_1(y(t_\alpha)) + (1-q_\alpha)u_1(y(t_\beta)). \end{aligned}$$

We also compute the following:

$$\begin{aligned} U_1(f_{k,K-1} \circ \gamma) &= q_\alpha u_1(a_k) + (1-q_\alpha)u_1(a_k) = u_1(a_k) \\ U_1(y \circ \gamma) &= q_\alpha u_1(y(t_\alpha)) + (1-q_\alpha)u_1(y(t_\alpha)) = u_1(y(t_\alpha)). \end{aligned}$$

Then, condition $(*)$ for agent 1 can be translated into the following condition:

$$q_\alpha \left[u_1(a_k) - u_1(y(t_\alpha))\right] + (1-q_\alpha)\left[u_1(a_{K-1}) - u_1(y(t_\beta))\right] \geq 0 \Rightarrow u_1(a_k) \geq u_1(y(t_\alpha)),$$

which of course holds since a_{K-1} is the bottom-ranked alternative for agent 1.

Hence, condition $(*)$ holds. We conclude that the SCS F violates Bayesian monotonicity. ■

Remark: If K is large, the example exhibits a severe failure of Bayesian monotonicity, to the extent that many SCFs in the SCS violate its requirement.

However, we have the following:

Fact 6 *The SCS F in the example of Section 4 satisfies uniform Bayesian monotonicity.*

Proof: Given Fact 1, this follows from Theorem 3. ■

As an illustration of the above fact, we offer the details of the argument for a specific deception. Consider a deception $\gamma_2(t_\alpha) = \gamma_2(t_\beta) = t_\alpha$ and an SCF $f_{2,K-1} \in F$. Since we have $f_{2,K-1}(\gamma_2(t_\alpha)) = f_{2,K-1}(\gamma_2(t_\beta)) = a_2$, it follows that $f_{2,K-1} \circ \gamma \notin F$. Consider a lottery y satisfying

$$(A)\ U_2(f_{2,K-1}|t_\alpha) \geq U_2(y|t_\alpha) \text{ and } U_2(f_{2,K-1}|t_\beta) \geq U_2(y|t_\beta).$$

Clearly, such a lottery y exists because $f_{2,K-1}$ specifies the second-best outcome in each state for agent 2, i.e., a_2 for type t_α and a_{K-1} for type t_β. For example, $y = a_3$ satisfies condition (A).

However, if condition $(**)$ in the definition of uniform Bayesian monotonicity were to hold, we must satisfy

$$(B)\ U_2(f_{2,K-1} \circ \gamma|t_\alpha) \geq U_2(y \circ \gamma|t_\alpha) \text{ and } U_2(f_{2,K-1} \circ \gamma|t_\beta) \geq U_2(y \circ \gamma|t_\beta),$$

which is not true for $y = a_3$ and type t_β. In other words, the change in preferences from state α to state β facilitates the existence of preference reversals, which make condition $(**)$ hard to meet.

6 Sufficient Conditions for Rationalizable Implementation in General Environments

In this section, we present a very general sufficiency result for full implementation in rationalizable strategies.

For each SCF $f \in \mathbb{F}$, define

$$Y_i[f] \equiv \left\{ y_i : T_{-i} \to \Delta^*(A) | U_i(f|\tilde{t}_i) \geq U_i(y_i|\tilde{t}_i)\ \forall \tilde{t}_i \in T_i^* \right\}.$$

The set $Y_i[f]$ describes the SCFs that are not better than the SCF f for agent i, regardless of his type.[9] Since both T_{-i} and $\Delta^*(A)$ are finite, the set $Y_i[f]$ becomes finite.[10]

Definition 9 *An SCS F satisfies the* ***no-worst-rule (NWR)*** *condition if, for each $f \in F$, $i \in N$, $t_i \in T_i^*$, and $\psi_i \in \Delta(T_{-i} \times T_{-i})$, there exist two SCFs*

[9] Here we adopt the notation from Oury and Tercieux (2012).

[10] In the appendix (Section A.3.3), we argue how to dispense with the finiteness of these sets.

$y_i[f;t_i,\psi_i], y_i^{'}[f;t_i,\psi_i] \in Y_i[f]$ *such that*

$$\sum_{t_{-i},t_{-i}^{'}} \psi_i(t_{-i}^{'},t_{-i})u_i(y_i^{'}[f;t_i,\psi_i](t_{-i}^{'});(t_i,t_{-i})) > \sum_{t_{-i},t_{-i}^{'}} \psi_i(t_{-i}^{'},t_{-i})u_i(y_i[f;t_i,\psi_i](t_{-i}^{'});(t_i,t_{-i})).$$

Using the example of Section 4, we confirm below that NWR is not necessary for implementation in rationalizable strategies.

Fact 7 *The SCS F in the example of Section 4 violates NWR.*

To see this, take the SCF $f_{K-1,K-1} \in F$, which assigns outcome a_{K-1} in both states. Since a_{K-1} is the worst outcome in both states for agent 1, the SCS F violates NWR. Nevertheless, by Fact 1, we know that the SCS F is implementable in rationalizable strategies.

Since T_i is finite, we define $\{t_i^\ell\}_{\ell=1}^\infty$ as a countable support of $\Delta(T_{-i})$. Similarly, since T_{-i} is countable, one can find $\{\psi_i^k\}_{k=1}^\infty$ as a countable support of $\Delta(T_{-i}\times T_{-i})$. More precisely, every ψ_i^k is defined as a degenerate measure over $T_{-i}\times T_{-i}$.

Since F satisfies NWR, for each $f \in F$ and $i \in N$, we define the uniform SCF $\bar{y}_i[f]$ as follows: there exist $\delta,\eta \in (0,1)$ such that:

$$\bar{y}_i[f] \equiv \frac{(1-\delta)(1-\eta)}{2}\sum_{\ell=1}^{\infty}\eta^{\ell-1}\sum_{k=1}^{\infty}\delta^{k-1}\left\{y_i^{'}[f;t_i^\ell,\psi_i^k]+y_i[f;t_i^\ell,\psi_i^k]\right\}.$$

We use this uniform SCF $\bar{y}_i[f]$ in the canonical mechanism proposed later.

We also use the following result later, an implication of NWR:

Lemma 1 *Suppose that an SCS F satisfies NWR. Then, for all $i \in N$, $t_i \in T_i^*$, and $\phi_i \in \Delta(T_{-i})$, there exist two lotteries (or constant SCFs) $\alpha_i[t_i,\phi_i], \alpha_i^{'}[t_i,\phi_i] \in \Delta^*(A)$ such that*

$$\sum_{t_{-i}} \phi_i(t_{-i})u_i(\alpha_i^{'}[t_i,\phi_i];(t_i,t_{-i})) > \sum_{t_{-i}} \phi_i(t_{-i})u_i(\alpha_i[t_i,\phi_i];(t_i,t_{-i})).$$

Proof: The proof can be found in Kunimoto (2019). ■

Following the previous argument, we denote by $\{t_i^\ell\}_{\ell=1}^\infty$ the countable support of T_i and by $\{\phi_i^k\}_{k=1}^\infty$ the countable support of T_{-i}, respectively. More precisely, each ϕ_i^k is a degenerate probability measure on T_{-i}. For each $i \in N$, we define the uniform lottery $\bar{\alpha}_i \in \Delta^*(A)$ as follows: there exist $\delta,\eta \in (0,1)$ such that

$$\bar{\alpha}_i \equiv \frac{(1-\delta)(1-\eta)}{2}\sum_{\ell=1}^{\infty}\eta^{\ell-1}\sum_{k=1}^{\infty}\delta^{k-1}\left\{\alpha_i^{'}[t_i^\ell,\phi_i^k]+\alpha_i[t_i^\ell,\phi_i^k]\right\}.$$

Finally, we define

$$\bar{\alpha} \equiv \frac{1}{n} \sum_{i \in N} \bar{\alpha}_i.$$

We also use this uniform lottery $\bar{\alpha}$ in the canonical mechanism.

The next two weak conditions are the last ones used in the general sufficiency theorem:

Definition 10 *An SCS F satisfies the **minimal conflict-of-interests (MCI)** condition if there do not exist $\beta \in \mathcal{B}$ and $f \in F \circ \beta$ such that $|F \circ \beta| \geq 2$ and $f \in \arg\max_{\tilde{f} \in F \circ \beta} U_i(\tilde{f}|t_i)$ for each $i \in N$ and $t_i \in T_i$.*

Definition 11 *An SCS F satisfies **responsiveness to deceptions (RD)** if there exists no deception $\beta \in \mathcal{B}$ such that $F \circ \beta \subseteq F$.*

Using the example of Section 4, we confirm below that MCI and RD are not necessary for implementation in rationalizable strategies.

Fact 8 *The SCS F in the example of Section 4 violates MCI. It also violates RD.*

To see that MCI is violated, consider a deception $\beta_2(t_\alpha) = \beta_2(t_\beta) = t_\alpha$ and the SCF $f_{K,K} \in F$, which assigns outcome a_K in both states. We observe that $|F \circ \beta| \geq 2$ and $f_{K,K} \in F$. However, since a_K is the best outcome for both agents in both states, we have $f_{K,K} \in \arg\max_{\tilde{f} \in F \circ \beta} U_i(\tilde{f}|t_i)$ for each $i \in N$ and $t_i \in T_i$, which contradicts MCI.

Next, to see that RD is violated, consider a deception $\beta_2(t_\alpha) = \beta_2(t_\beta) = t_\alpha$ and the SCF $f_{1,K-1} \in F$. Then, we have that $f_{1,K-1} \circ \beta = f_{1,1}$, which assigns outcome a_1 in both states. However, $f_{1,1} \notin F$, which contradicts RD.

Assume that the SCS F satisfies (fully) rationalizable incentive compatibility (fully RIC). Then, there exist an expanded type space $\mathcal{T} = (\mathcal{T}_i, T_i, \hat{t}_i, \hat{\pi}_i)_{i \in N}$ and an extended SCF $\tilde{f} : \mathcal{T} \to \Delta(A)$ such that the identity mapping $I : \mathcal{T} \to \mathcal{T}$ constitutes a rationalizable profile in the associated direct mechanism $\Gamma^{DR}(\mathcal{T}, \tilde{f})$ and $\tilde{f} \circ \hat{t}^{-1} \approx F$.

For the sufficiency result we establish below, we propose the following mechanism $\Gamma = (M, g)$: each agent i sends a message $m_i = (m_i^1, m_i^2, m_i^3, m_i^4, m_i^5, m_i^6)$, where

- $m_i^1 \in \mathcal{T}_i$, an extended type for agent i;
- $m_i^2 \in F$, i.e., a social choice function, understood as a recommendation to the designer;

- $m_i^3 = (m_i^3[1], m_i^3[2])$ where $m_i^3[1] : T_{-i} \to \Delta^*(A)$ and $m_i^3[2] \in F$, understood as potential arguments for a challenge to the designer;
- $m_i^4 \in \Delta^*(A)$, i.e., a state-independent allocation, also understood as a challenge to the designer;
- $m_i^5 \in N$, i.e., a number chosen from $\{1, \ldots, n\}$, understood as a vote for some person to be the king;
- and $m_i^6 \in \mathbb{N}$, i.e., a positive integer.

The outcome function $g : M \to \Delta(A)$ is defined as follows: for each $m \in M$:

Rule 1. Consensus implements the recommendation made by the elected king: If $m_i^6 = 1$ for all $i \in N$, then $g(m) = f(\hat{t}(m^1))$, where $f = m_k^2$ and $k = (\sum_{j \in N} m_j^5) \pmod{n+1}$.

Rule 2. An odd man out: If $m_j^6 = 1$ for all $j \neq i$ and $m_i^6 > 1$, then the following subrules apply:

Rule 2-1. A nongreedy odd man out is heard in his challenge, although some bad outcomes –cost of challenge– are also implemented in the appeal process: If $U_i(m_k^2|\tilde{t}_i) \geq U_i(m_i^3[1]|\tilde{t}_i)$ for all $\tilde{t}_i \in T_i$ and $m_k^2 = m_i^3[2]$ where $k = (\sum_{j \in N} m_j^5) \pmod{n+1}$, then

$$g(m) = \begin{cases} m_i^3[1](\hat{t}_{-i}(m_{-i}^1)) & \text{with probability } m_i^6/(m_i^6+1) \\ \bar{y}_i[m_k^2](\hat{t}_{-i}(m_{-i}^1)) & \text{with probability } 1/(m_i^6+1) \end{cases}$$

Rule 2-2. A greedy odd man out is not heard in his challenge, although some bad outcomes –cost of the challenge– are also implemented in the appeal process: Otherwise,

$$g(m) = \begin{cases} m_k^2(\hat{t}(m^1)) & \text{with probability } m_i^6/(m_i^6+1) \\ \bar{y}_i[m_k^2](\hat{t}_{-i}(m_{-i}^1)) & \text{with probability } 1/(m_i^6+1) \end{cases}$$

where $k = (\sum_{j \in N} m_j^5) \pmod{n+1}$.

Rule 3. Stronger disagreements lead to an integer game, implementing

potential disarray in the appeal/challenge process: In all other cases,

$$g(m) = \begin{cases} m_1^4 & \text{with probability } \frac{m_1^6}{n(m_1^6+1)} \\ m_2^4 & \text{with probability } \frac{m_2^6}{n(m_2^6+1)} \\ \vdots & \vdots \\ m_n^4 & \text{with probability } \frac{m_n^6}{n(m_n^6+1)} \\ \bar{\alpha} & \text{with the remaining probability,} \end{cases}$$

We are finally ready to state the general sufficiency result for full implementation in rationalizable strategies:

Theorem 4 *If an SCS F satisfies fully rationalizable incentive compatibility, uniform Bayesian monotonicity, closure, NWR, MCI, and RD, it is **fully** implementable in rationalizable strategies.*

Proof: The proof is in the appendix (Section A.1). ■

A simple sketch of the proof is this. The proof is based on four steps. Step 1 uses the integer game in Rule 3 and NWR to show that, in all rationalizable messages, the integer announced in m_i^6 must be 1. Basically, Rules 2 or 3 cannot happen with positive probability, because if they did, an agent would find a better response by announcing a higher integer than he initially announces. Step 2 follows then easily, stating that any belief used to support rationalizable messages must put probability 1 on Rule 1. Step 3 uses (fully) RIC to establish that, for every $f \in F$, there exist rationalizable messages whose induced SCF is f. Finally, in Step 4, RD and UBM are used to obtain a key preference reversal, for any deception β – this preference reversal is obtained for free in the economic environments of Section 7. The preference reversal identifies a test-agent and a test-allocation that he would like to impose. Such a reversal, along with MCI, allows us to construct a profitable deviation from Rule 1, by allowing the test-agent with the reversal to impose the test-allocation that benefits him. This last argument establishes that any rationalizable outcome must be in the SCS F, which concludes the proof.

Remark: Theorem 4 also applies to environments with two agents. In contrast, our result for complete-information environments, in Kunimoto and Serrano (2019), is proved for the case of at least three agents, because, as is usual in those settings, we construct a mechanism that relies on the report of the entire state. With RIC, we get around that issue here.

7 Economic Environments

In this section, we propose a class of well-studied economic environments in which all the additional conditions used in Theorem 4, such as UBM, NWR, MCI, and RD can be dispensed with. It follows that, in this class of environments, RIC is essentially the only relevant condition for rationalizable implementation.

7.1 First Price Auctions under Independent Private Values

Although the result in this section can be extended to many other economic environments, to fix ideas, we work with a well-known auction setting. We define an environment that is tied to a specific trading institution (in this case, an independent private-values first-price auction for two bidders – see, e.g., Battigalli and Siniscalchi (2003)). Let $N = \{1, 2\}$ be the set of agents. Let $T_i = [0, 1]$ for $i = 1, 2$ be the set of agent i's types or valuations, drawn independently from a continuous probability distribution. Let A denote the set of outcomes, defined as follows:

$$A = \{(1, p), (0, 0)\}_{p\in[0,1]} \cup \{(0, 0), (1, p)\}_{p\in[0,1]}$$

such that there exist bidding strategies $\sigma_i : T_i \mapsto [0, 1]$, $i = 1, 2$, so that agent i of type t_i gets the good with probability one if and only if $p = \sigma_i(t_i) > \sigma_j(t_j)$ and pays a price p. That is, either agent 1 gets the object paying potentially any price $p \in [0, 1]$, while agent 2 does not get the object and pays nothing, or vice versa, always obeying the rules of the first-price auction. Of course, the set A is endowed with its sigma-algebra $\mathscr{A}$ containing all singleton sets. Let $\Delta(A)$ be the set of probability measures over this measurable space.

In this economic environment, under a very weak regularity assumption that is satisfied by standard bidding strategies, we can establish the following characterization of SCSs that are implementable in rationalizable strategies:

Theorem 5 *Suppose that an SCS F contains a continuous SCF $f \in F$ such that for every agent $i = 1, 2$, (i) $U_i(f|0) = 0$, and (ii) for $t_i \in (0, 1]$, $U_i(f|t_i) > 0$. Then, the SCS F is implementable in rationalizable strategies if and only if it satisfies rationalizable incentive compatibility and closure.*

Proof of Theorem 5: The proof is in the appendix (Section A.2). ■

We provide a sketch of the proof. The necessity of RIC and closure has been established in Theorems 1 and 2, respectively. Sufficiency follows from the general sufficiency theorem, i.e., Theorem 4, found in Section 6 and extended in Section A.3 in the appendix to environments where type spaces are compact subsets of Euclidean spaces, after taking into account the following observations. In economic

environments, as usual, the *minimal conflict of interest* (MCI) condition is trivially satisfied: after the assumed existence of the SCF f, if the SCS F contains additional SCFs, there is no SCF that can simultaneously be the maximizer of every type and every agent's expected utility within F, given the definition of the set of alternatives in the first-price auction. Furthermore, the *no worst rule* (NWR) condition, used to impose bad outcomes after deviations, can be dispensed with in economic environments as well, by always using the "Zero" outcome, i.e., an outcome where the good is allocated to the deviating agent with probability zero. The heart of the proof of Theorem 5 lies in Proposition 4, which is presented and proved in Section A.1. This proposition establishes that, in these environments, after any deception that the agents might use, there always exists a test-agent and a test-pair of SCFs that can help the designer circumvent the deception. That is, around the assumed f, there exists an agent i, a type t_i, and an SCF y such that type t_i prefers f over y if agents are not using any deception, but has the opposite preference when the deception is used. This fundamental fact for these environments implies that *uniform Bayesian monotonicity* (UBM), an additional necessary condition for full implementability in rationalizable strategies, is trivially satisfied. Moreover, the existence of that preference reversal also lets one dispense with the *responsiveness to deceptions* (RD) condition, which is the last assumption in Theorem 4. It follows then from Theorem 4 that *rationalizable incentive compatibility* (RIC) and *closure* are sufficient for implementability in rationalizable strategies.

In these environments, the type space $T = \prod_{i \in N} T_i$ is belief-closed, and hence, closure can be obviated. We conclude, therefore, that implementability in rationalizable strategies is equivalent to RIC.

To illustrate our analysis, we revisit Battigalli and Siniscalchi (2003, Section 2), who identify the set of rationalizable bidding strategies in the independent, private-values, first-price auction. For example, for the case of the uniform probability distribution on $[0, 1]$, the set of rationalizable strategies consists of: $\sigma_i(0) = 0$ and $\sigma_i(t_i) \in (0, t_i/2]$ for $t_i \in (0, 1]$. (Recall that the Bayesian equilibrium consists of $\sigma_i(t_i) = t_i/2$ for $i = 1, 2$ and for all t_i in this case.) This identifies a set of SCFs, each of which is associated with each agent using one of the bidding strategies in this rationalizable set. Now, it follows from Theorem 5, and from the fact that closure can be obviated, that the set of SCFs identified by Battigalli and Siniscalchi (2003) is the maximal SCSs that can be implemented in rationalizable strategies, whatever the mechanism one uses. To see this, note that, if one had an SCF outside of the Battigalli-Siniscalchi set that is also supported by rationalizable strategies in some mechanism, given the way we have defined the set A of outcomes, by RIC, the "revelation principle" embodied in our Theorem 1, such an SCF would have to be rationalizable in the first-price auction, contrary to what we are assuming.

7.2 Multidimensional Signals

Following Krishna and Perry (2000), one can use the same proof in order to get a sense of the large class of economic environments for which we can establish Theorem 5. These are not necessarily tied to a specific trading institution. Indeed, there is a finite set K of social alternatives. Each agent i then has a K-dimensional type $t_i = (t_i(1), t_i(2), \ldots, t_i(K)) \in \mathbb{R}^K$. The payoff to agent i of type t_i for alternative k is quasilinear, that is, it is of the form: $t_i(k) - x_i$, where x_i is a monetary transfer made by i to the planner. The environment to be considered is thus one of private values. We assume that for all $i \in N$, T_i is a nonempty compact and convex subset of $\mathbb{R}^K$. We also assume the existence of a type $\underline{t}_i \in T_i$ for each $i \in N$ such that $\underline{t}_i$ is the "most reluctant" type of agent i in the sense that his gain from participating in the mechanism is the least among all types of agent i. We call $\underline{t}_i$ the *zero* type for agent i. Then, each SCS F to be considered contains an SCF f for which the zero type of each agent i obtains zero expected utility ($U_i(f|\underline{t}_i) = 0$) and every nonzero type obtains positive expected utility ($U_i(f|t_i) > 0$ for every $t_i \neq \underline{t}_i$).

Moving beyond private values, Jehiel and Moldovanu (2001) presents a remarkable impossibility result for environments with multidimensional signals. Indeed, in their settings, no efficient rule is Bayesian incentive compatible. We can illustrate our approach by constructing a simple mechanism to show that efficient rules can be part of SCSs satisfying RIC, allowing their implementation in rationalizable strategies. We describe the elements of Example 4.4 in their paper, on which we base our construction.

There are two agents $i = 1, 2$ and three alternatives $k = A, B, C$. Suppose that only agent 1 receives a signal, denoted by $s = (s_A, s_B, s_C)$. Let S denote the set of signals, assumed to be a compact and convex subset of $\mathbb{R}^3_+$.

Assume the set of agents' valuations is such that alternative C is never efficient, say $v_1(C, s) = v_2(C, s) = 0$ for all $s \in S$, whereas for all $s \in S$, $v_i(A, s) \geq 0$ and $v_i(B, s) \geq 0$ for $i = 1, 2$. Moreover, we have that $v_i(A, s) = v_i(B, s) = 0$ if and only if $s = (0, 0, 0)$ for $i = 1, 2$. So, excluding the trivial signal $s = (0, 0, 0)$, either only alternative A is efficient, only alternative B is efficient, or both alternatives A and B are efficient.

Denote by $S_A \subseteq S$ the set of signals for which A is the only efficient alternative, i.e.,

$$S_A = \{s \in S : v_1(A, s) + v_2(A, s) > v_1(B, s) + v_2(B, s)\}.$$

Similarly, let $S_B \subseteq S$ be the set of signals for which B is the only efficient alternative, i.e.,

$$S_B = \{s \in S : v_1(B, s) + v_2(B, s) > v_1(A, s) + v_2(A, s)\}.$$

And let $S_{AB} = S \setminus (S_A \cup S_B)$, where both A and B are efficient. This third set has zero measure in S.

We only care about agent 1's misrepresentation of her type to the extent that it induces a distinct alternative which is not efficient under a true signal profile. That is, if we were concerned with allocative efficiency, we would like an SCS F containing only SCFs f such that $f(s) = A$ – and perhaps some monetary transfers – whenever $s \in S_A$, and $f(s) = B$ – and perhaps some monetary transfers– whenever $s \in S_B$. Over the set S_{AB}, the SCF f could assign either alternative A or B, again with some transfers.

Consider the following mechanism. Agent 1's message set is the set of her signals S. Agent 2 can "veto" (V) or "not veto" (NV) the interaction. The outcome function is as follows:

- If agent 2 chooses NV, we listen to the report of agent 1, say $\hat{s}$. Then, (i) the outcome is alternative A if $\hat{s} \in S_A \cup S_{AB}$ and agent 2 makes a transfer $v_2(A, \hat{s})$ to agent 1; and (ii) the outcome is alternative B if $\hat{s} \in S_B$ and agent 2 makes a transfer $v_2(B, \hat{s})$ to agent 1.

- If agent 2 chooses V, alternative C is implemented and there are no transfers.

It is easy to see that, in this mechanism, all messages are rationalizable, as we argue next. If agent 2 believes that agent 1 is not misrepresenting her signal, this gives agent 2 an ex-post zero payoff against the type who receives that signal. Thus, if agent 2 believes that agent 1's strategy is truth-telling for all signals, his expected payoff from the NV action is zero. Clearly, his expected payoff from choosing his V action is also zero. Thus, he is at a best response by choosing either message. For agent 1, any strategy is a best response if she believes that agent 2 is choosing V with probability 1. Thus, truth-telling as well as any deception can be rationalized by agent 1. These arguments already establish that the entire interim efficient frontier may arise as a result of rationalizable play in this mechanism, although there are also inefficient outcomes that are part of the rationalizable set.

Suppose $s \in S_A \cup S_{AB}$. Let $D_-(s)$ be the following set of reports $\hat{s}$: $\hat{s} \in D_-(s)$ whenever $\hat{s} \in S_A \cup S_{AB}$ and $v_2(A, \hat{s}) < v_2(A, s)$. Similarly, let $D_+(s)$ be the set of reports $\hat{s} \in S_A$ such that $v_2(A, \hat{s}) \geq v_2(A, s)$. Also, let $D'_-(s)$ be the set of reports $\hat{s} \in S_B$ such that

$$v_1(B, s) + v_2(B, \hat{s}) < v_1(A, s) + v_2(A, s).$$

and let $D'_+(s)$ be the set of reports $\hat{s} \in S_B$ such that

$$v_1(B, s) + v_2(B, \hat{s}) \geq v_1(A, s) + v_2(A, s).$$

Note how the outcome of truth-telling is the efficient rule that maximizes 1's payoff leaving a zero payoff to agent 2. But in addition, any efficient rule can

arise in rationalizable play when agent 1 uses deceptions in the set $D_{-}(s)$, and hence transfering some of the available surplus to agent 2. Deceptions that utilize the other sets just defined, such as $D_{+}(s), D'_{-}(s)$, and $D'_{+}(x)$, will result in inefficiencies.

This "anything goes" mechanism just starts to scratch the surface of possibilities, in terms of implementing efficiency as the outcome of rationalizable play. There will surely be other mechanisms that deliver a more refined outcome, perhaps getting rid of all inefficiencies; they are beyond our current scope.

8 The Case of Social Choice Functions

In this section, we confine our attention to SCFs and investigate the implications of their rationalizable implementation. In the process, we clarify some results in the literature. First, we recall the definition of Bayesian monotonicity, particularized for SCFs, but stated in its contrapositive form, in order to enhance the comparison with the next condition, presented hereafter:

Definition 12 *An SCF f satisfies **Bayesian monotonicity** (BM) if, for every deception β, whenever $f \circ \beta \not\approx f$, there exist $i \in N$, $t_i \in T_i^*$, and $y^* : T_{-i} \to \Delta(A)$ such that:*

$$\left[\begin{array}{ll} U_i(y^* \circ \beta | t_i) > U_i(f \circ \beta | t_i) & \\ and \qquad U_i(f|\tilde{t}_i) \geq U_i(y|\tilde{t}_i) & \forall \tilde{t}_i \in T_i \end{array} \right].$$

Oury and Tercieux (2012) introduces the notion of *interim rationalizable monotonicity*, and shows it to be an implication of continuous implementation in their setting. To define this condition, let $\beta_i : T_i \to 2^{T_i} \backslash \{\emptyset\}$ be a set-valued deception of agent i and we call $\beta = (\beta_1, \ldots, \beta_n)$ a set-valued deception. We say $f \circ \beta \approx f$ if, for all $t, t' \in T^*$, whenever $t' \in \beta(t)$, $f(t) = f(t')$. Otherwise, we say $f \circ \beta \not\approx f$, and we refer to such deceptions as *unacceptable deceptions*. Then, we introduce the following condition.

Definition 13 (Oury and Tercieux (2012)) *An SCF f satisfies **interim rationalizable monotonicity** (IRM) if, for every set-valued deception β for which $f \circ \beta \not\approx f$, there exist $i \in N$, $t_i \in T_i^*$, and $t'_i \in \beta_i(t_i)$ such that for every $\psi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying*

1. $\psi_i(t_{-i}, t'_{-i}) > 0 \Rightarrow t'_{-i} \in \beta_{-i}(t_{-i})$,
2. $\sum_{t'_{-i}} \psi_i(t_{-i}, t'_{-i}) = \pi_i(t_i)[t_{-i}]$,

there exists an SCF $y^* : T_{-i} \to \Delta(A)$ *such that*

$$\sum_{t_{-i},t'_{-i}} \psi_i(t_{-i}, t'_{-i}) u_i(y^*(t'_{-i}); (t_i, t_{-i})) > \sum_{t_{-i},t'_{-i}} \psi_i(t_{-i}, t'_{-i}) u_i(f(t'_i, t'_{-i}); (t_i, t_{-i}));$$

and for all $\tilde{t}_i \in T_i^*$,

$$\sum_{t_{-i}} \pi_i(\tilde{t}_i)[t_{-i}] u_i(f(\tilde{t}_i, t_{-i}); (\tilde{t}_i, t_{-i})) \geq \sum_{t_{-i}} \pi_i(\tilde{t}_i)[t_{-i}] u_i(y^*(t_{-i}); (\tilde{t}_i, t_{-i})).$$

The following results can be found also in Oury and Tercieux (2012):

Proposition 2 *If an SCF* f *satisfies interim rationalizable monotonicity, it also satisfies SIRBIC, and hence, also Bayesian incentive compatibility.*

Proof: It follows from Lemma 3 of Oury and Tercieux (2012). ■

Proposition 3 *If an SCF* f *satisfies interim rationalizable monotonicity, it also satisfies Bayesian monotonicity.*

Proof:[11] Consider a single-valued deception β and suppose that there exists a state $t \in T^*$ such that $f \circ \beta(t) \neq f(t)$. Hence, β is such that $f \circ \beta \not\approx f$. Then, notice that, by IRM, using the fact that β is single-valued, we can get rid of the mapping ψ_i, and conclude that there exists an SCF $y^* : T_{-i} \to \Delta(A)$ such that

$$\sum_{t_{-i}} \pi_i(t_{-i}|t_i) u_i(y^*(\beta_{-i}(t_{-i}); (t_i, t_{-i})) > \sum_{t_{-i}} \pi_i(t_{-i}|t_i) u_i(f(\beta(t); (t_i, t_{-i}));$$

and for all $\tilde{t}_i \in T_i^*$,

$$\sum_{t_{-i}} \pi_i(\tilde{t}_i)[t_{-i}] u_i(f(\tilde{t}_i, t_{-i}); (\tilde{t}_i, t_{-i})) \geq \sum_{t_{-i}} \pi_i(\tilde{t}_i)[t_{-i}] u_i(y^*(t_{-i}); (\tilde{t}_i, t_{-i})).$$

Hence, we have found the necessary preference reversal, as specified in the requirement of Bayesian monotonicity. ■

Notice that Bayesian monotonicity does not necessarily imply interim rationalizable monotonicity. To understand the reason, let us consider an SCF satisfying BM. To prove that it satisfies IRM, consider an unacceptable set-valued deception β. There exists a single-valued selection $\hat{\beta}$ of β which is also unacceptable. So indeed this single-valued selection $\hat{\beta}$ can be undermined as required in IRM using

[11] To better follow the development of the examples in this section, it is useful to provide this proof here.

BM. That is, there exist i, t_i, and $t_i' = \hat{\beta}_i(t_i)$ for which the requirements of IRM are satisfied for the *unique belief* $\hat{\psi}_i$ that is consistent with the single-valued deception $\hat{\beta}_{-i}$. But this is not enough to argue that i, t_i, and $t_i' \in \beta_i(t_i)$ satisfy the requirements of IRM for *all beliefs* ψ_i that are consistent with the set-valued β_{-i}. This may end up being too strong a requirement.

And indeed, the next example makes exactly this point:

Example 1 *The following SCF f satisfies BM, but violates IRM. We write down the example starting to describe only one agent called agent 1, who is informed of the state, but we add an uninformed agent when needed to complete the argument.*

There are two states $T = \{t, t'\}$. There are four pure outcomes $A = \{a, b, c, d\}$. Let $f(t) = a$ and $f(t') = b$. The Bernoulli utility function of the informed agent, whom we call agent 1, is given as follows:

$$\begin{aligned} u(a,t) &= 3; u(b,t) = 1; u(c,t) = 2; u(d,t) = 4; \\ u(a,t') &= 3; u(b,t') = 2; u(c,t') = 1; u(d,t') = 4. \end{aligned}$$

Note that the SCF f violates BIC (which coincides with RIC for SCFs), and hence, it is not implementable either in rationalizable strategies or in Bayesian equilibrium. That the SCF f fails BIC implies that it also violates IRM, which is implied by Proposition 2 above (Lemma 3 in Oury and Tercieux (2012)). But we find it instructive, in this and the next example, to show this violation explicitly. Also, as will be shown, the SCF f satisfies BM. Serrano and Vohra (2001) shows by means of examples that there are many nonconstant SCFs satisfying BIC, while only constant SCFs satisfy BM. Hence, we can conclude that BIC and BM are logically independent.

Let us thus show that the SCF f satisfies BM. Recall that it suffices that the SCF y^* chosen in the prerequisite for BM be constant in this case. There are three single-valued deceptions to consider:

- *$\beta_1(t) = t'$ and $\beta_1(t') = t$. Then, the test-agent is agent 1, the test-type is t and the SCF creating a reversal is $y^* = c$. To see this, we display the following expected utilities of agent 1:*

$$\begin{aligned} U_1(f \circ \beta_1 | t) &= u(b,t) = 1 < 2 = u(c,t) = U_1(y^* \circ \beta_1 | t) \\ U_1(f|t) &= u(a,t) = 3 > 2 = u(c,t) = U_1(y^*|t) \\ U_1(f|t') &= u(b,t') = 2 > 1 = u(c,t') = U_1(y^*|t') \end{aligned}$$

- *$\beta_2(t) = \beta_2(t') = t$. Add agent 2 with a Bernoulli utility function v, who is uninformed about the state, and for whom her expected utilities, calculated using a uniform probability distribution over the states, are based on $v(a,t) =$*

1, $v(b,t^{'})=3$ *–which are the Bernoulli utilities of the nonmanipulated SCF–, completed with* $v(a,t^{'})=1$*, and* $v(d,t)=v(d,t^{'})=1.5$*. The other utility values are not needed for now. Then, the test-agent is agent 2 and the SCF* $y^{*}=d$*. To see this, we display the following expected utilities of agent 2:*

$$\begin{aligned} U_2(f\circ\beta_2) &= \frac{1}{2}v(a,t)+\frac{1}{2}v(a,t^{'})=1<1.5=\frac{1}{2}v(d,t)+\frac{1}{2}v(d,t^{'})=U_2(y^{*}\circ\beta_2) \\ U_2(f) &= \frac{1}{2}v(a,t)+\frac{1}{2}v(b,t^{'})=2>1.5=\frac{1}{2}v(d,t)+\frac{1}{2}v(d,t^{'})=U_2(y^{*}) \end{aligned}$$

- $\beta_3(t)=\beta_3(t^{'})=t^{'}$*. Now, the test-agent is again agent* 1*, the test-type is* t*, and the test-SCF* $y^{*}=c$*. To see this, we display the following expected utilities of agent 1:*

$$\begin{aligned} U_1(f\circ\beta_3|t) &= u(b,t)=1<2=u(c,t)=U_1(y^{*}\circ\beta_3|t) \\ U_1(f|t) &= u(a,t)=3>2=u(c,t)=U_1(y^{*}|t) \\ U_1(f|t^{'}) &= u(b,t^{'})=2>1=u(c,t^{'})=U_1(y^{*}|t^{'}) \end{aligned}$$

Hence, for each single-valued deception, we have found the preference reversal required by BM. Thus, BM is satisfied.

To argue that the SCF f *violates IRM, consider the following set-valued deception* $\beta=\{\beta_1,\beta_2\}$*:* $\beta_1(t)=\{t\}$ *and* $\beta_2(t^{'})=\{t,t^{'}\}$*. This set-valued deception cannot be undermined as per the requirements of IRM. Agent 2 does not have the required preference reversal for* all beliefs ψ_2 *that are consistent with this deception because one such belief is that agent 1 is truthful, i.e.,* $\psi_2(t,t)=\psi_2(t^{'},t^{'})=1/2$*, and hence, it is impossible to have* $U_2(y)>U_2(f)\geq U_2(y)$*. Agent 1 can hold only one possible belief about agent 2, that is, agent 2 is truth-telling since he has a unique type. Now, consider type* t *of agent 1 and* $t\in\beta(t)$*. Clearly, there does not exist any lottery* $y\in\Delta(\{a,b,c,d\})$ *such that* $U_1(y|t)=u(y,t)>u(f(t)|t)=U_1(f|t)$ *and* $U_1(f|t)=u(f(t),t)\geq u(y,t)=U_1(y|t)$*. Next, consider type* $t^{'}$ *of agent 1 and* $t^{'}\in\beta(t^{'})$*. Again, there does not exist any lottery* y *such that* $U_1(y|t^{'})=u(y,t^{'})>u(f(t^{'}),t^{'})=U_1(f|t^{'})$ *and* $U_1(f|t^{'})=u(f(t^{'})|t^{'})\geq u(y|t^{'})=U_1(y|t^{'})$*. Finally, consider type* $t^{'}$ *of agent 1 and* $t\in\beta(t^{'})$*. There does not exist any lottery* y *such that* $U_1(y|t^{'})=u(y,t^{'})>u(f(t),t^{'})=U_1(f|t^{'})$ *and* $U_1(f|t^{'})=u(f(t^{'}),t^{'})\geq u(y,t^{'})=U_1(y|t^{'})$*. Thus, the unacceptable set-valued deception* β *cannot be undermined, and IRM is violated.*

Our next task is to understand whether IRM is equivalent to BIC and BM imposed together. We already know, by Propositions 2 and 3, that IRM implies BIC and BM. However, the next example, a variant of Example 1, demonstrates that the opposite implication is not true:

Example 2 *The following SCF f satisfies BIC and BM, but violates IRM.*

There are two states t and $t^{'}$. There are four pure outcomes $\{a, b, c, d\}$. Let $f(t) = a$ and $f(t^{'}) = b$.

There are two agents, agent 1 with a Bernoulli utility function u and agent 2 with v, respectively. Agent 1 is informed, and agent 2 is uninformed, assigning equal probability to either state.

The utility function u of the informed agent (agent 1) is as follows:

$$u(a,t) = 3; u(b,t) = 1; u(c,t) = 2; u(d,t) = 4;$$
$$u(a,t^{'}) = 3; u(b,t^{'}) = 3; u(c,t^{'}) = 1; u(d,t^{'}) = -1.$$

Note that f now satisfies BIC (recall that BIC coincides with RIC for SCFs).

The Bernoulli utilities of the uninformed agent (agent 2) are denoted by v. They are identical to the ones in Example 1:

$$v(a,t) = 1; v(b,t) = 0; v(c,t) = 0; v(d,t) = 1.5;$$
$$v(a,t^{'}) = 1; v(b,t^{'}) = 3; v(c,t^{'}) = 0; v(d,t^{'}) = 1.5.$$

Similarly to Example 1, let us first check that f satisfies BM. Again, recall that it suffices that the proposed SCF y^* in the BM condition be constant in this case. There are three single-valued deceptions to consider:

- *$\beta_1(t) = t^{'}$ and $\beta_1(t^{'}) = t$. Then, the test-agent is 1, the test-type is t and the SCF creating a reversal is $y^* = c$. To see this, we display the following utilities:*

$$\begin{aligned} U_1(f \circ \beta_1|t) &= u(b,t) = 1 < 2 = u(c,t) = U_1(y^* \circ \beta_1|t) \\ U_1(f|t) &= u(a,t) = 3 > 2 = u(c,t) = U_1(y^*|t) \\ U_1(f|t^{'}) &= u(b,t^{'}) = 3 > 1 = u(c,t^{'}) = U_1(y^*|t^{'}). \end{aligned}$$

- *$\beta_2(t) = \beta_2(t^{'}) = t$. Then, the test-agent is agent 2 and the SCF $y^* = d$. We omit the computations, as they are identical to the ones in the previous example.*

- *$\beta_3(t) = \beta_3(t^{'}) = t^{'}$. Now, the test-agent is again agent 1, the test-type is t, and the test-SCF $y^* = c$. To see this, we display the following utilities:*

$$\begin{aligned} U_1(f \circ \beta_3|t) &= u(b,t) = 1 < 2 = u(c,t) = U_1(y^* \circ \beta_3|t) \\ U_1(f|t) &= u(a,t) = 3 > 2 = u(c,t) = U_1(y^*|t) \\ U_1(f|t^{'}) &= u(b,t^{'}) = 3 > 1 = u(c,t^{'}) = U_1(y^*|t^{'}). \end{aligned}$$

Thus, as in the previous example, for each single-valued deception, we have found the preference reversal required by BM. Therefore, BM is satisfied.

Also exactly as in Example 1, one can argue that the SCF f violates IRM, using again the same unacceptable set-valued deception.

Example 2 is important, as it shows that IRM is not a necessary condition for rationalizable implementation.[12] Indeed, the SCF in the example satisfies BIC (equivalent to RIC for SCFs), BM (equivalent to UBM for SCFs), closure, MCI, NWR (because of alternative c), and RD.[13] By our main sufficiency theorem (Theorem 4), the SCF in the example is implementable in rationalizable strategies.

It is possible, of course, that a simpler mechanism can be constructed for this specific example. Although we have not found it, we come close to doing that. The next mechanism approximately implements the SCF f in Example 2 in rationalizable strategies (or in (strict) Bayesian equilibrium). Let agent 1's message set be $M_1 = \{t, t'\}$. Let $M_2 = \{(T, L), (T, R), (B, L), (B, R)\}$, where T stands for "Top," B for "Bottom," L for "Left," and R for "Right." The outcome function is described in the following two tables; agent 1 chooses over tables, and agent 2 chooses over the four cells in the table (choosing, of course, the same cell across both tables). Let $\epsilon > 0$ be arbitrarily small:

Agent 1 chooses t		Agent 2	
		L	R
Agent 2	T	d	$(1-\epsilon)a + \epsilon d$
	B	d	d

Agent 1 chooses t'		Agent 2	
		L	R
Agent 2	T	c	b
	B	c	c

Note how it is a strictly dominant strategy for each of the informed types of agent 1 to announce the truthful message. Given that, in the second round of elimination, agent 2's unique best response is (T, R), and the result is the approximate implementation of the SCF f – exact in state t' – in rationalizable strategies.

[12] This contradicts an assertion made in Oury and Tercieux (2012, footnote 4), which also attributes this result to an incomplete-information adaptation of the types of argument used in Bergemann and Morris (2011) and Bergemann, Morris, and Tercieux (2011).

[13] The reader is referred to Example 2 of Kunimoto (2019) to see how to explicitly check that NWR is satisfied.

Here, the set of rationalizable strategies is a singleton, so that it corresponds to the unique strict Bayesian equilibrium. Defining the perturbed SCF f^ϵ as follows: $f^\epsilon(t) = (1-\epsilon)a + \epsilon d$ and $f^\epsilon(t') = b$, this implies that the SCF f^ϵ is exactly implementable in rationalizable strategies by the finite mechanism we constructed and the set of rationalizable strategy profiles is a singleton. Therefore, given the result by Dekel, Fudenberg, and Morris (2007), that the correspondence of (interim correlated) rationalizable strategies in a finite game is upper hemicontinuous in the product topology in the universal type space, f^ϵ is also strictly continuously implementable by the same finite mechanism.[14] By Theorem 3 of Oury and Tercieux (2012), we thus conclude that f^ϵ satisfies IRM, whereas the unperturbed SCF f violates it. In fact, we can also show that f^ϵ satisfies IRM explicitly.

Moreover, the mechanism when $\epsilon = 0$ fully and exactly implements the SCF f in Bayesian equilibrium, although the equilibrium ceases to be strict. However, the mechanism when $\epsilon = 0$ fails to implement f in rationalizable strategies, because, while truth-telling continues to be strictly dominant for type t, it is only weakly dominant for type t', even though the nontruthful report for type t' is a best response to one belief only, namely, that agent 2 chooses the cell (T, R) with probability one. This suggests that the insistence on "strict" continuous implementation, rather than continuous implementation, is crucial for Theorem 3 of Oury and Tercieux (2012). The very same point is also made by Chen, Kunimoto, and Sun (2019), who characterize (not strict) continuous implementation by allowing for small transfers.

To understand how our canonical mechanism succeeds in this example, the unacceptable deception used by agent 1, in which type t' pretends to be type t, is undermined by agent 2, who becomes the test-agent and induces Rule 2, imposing alternative d with arbitrarily high probability. We conjecture that this should still be feasible in a simpler mechanism, although it is not possible if one insists on finiteness (Bergemann and Morris (2008, Proposition 2)).

9 Conclusion

This paper has uncovered rationalizable incentive compatibility (RIC) of SCSs as the basis of a more permissive theory of incentives. Aside from RIC, closure and uniform Bayesian monotonicity have also been shown to be necessary for rationalizable implementation. Furthermore, all three are also sufficient, if one adds a few extra weak regularity conditions. Exploring other implications of RIC, for instance, revisiting key definitions under incomplete information, such as efficiency

[14] Proposition 1 of Dekel, Fudenberg, and Morris (2007) shows that all types that have the same hierarchies of beliefs have the same set of interim correlated rationalizable strategies.

(Holmström and Myerson (1983)) or the core (Wilson (1978)) should be part of our future research agenda.

A. Appendix

In this appendix, we provide all the omitted proofs of results in the paper. We also discuss how to extend our results to more general environments.

A.1. Proof of Theorem 4

By closure, without loss of generality, we restrict attention to type spaces where the event T is belief-closed. Also, for much of the proof, and in order to simplify notation, we work with the type t_i, as opposed to the extended type τ_i that we get from RIC. We only use the extended type τ_i in Step 3, where RIC is actually invoked.

We use the mechanism $\Gamma = (M, g)$ constructed in the main text. Now, we proceed to the formal proof.

Step 1: $\sigma_i(t_i) \in S_i^{\Gamma(T)}(t_i) \Rightarrow \sigma_i^6(t_i) = 1$.

Proof of Step 1: Fix $t_i \in T_i$. Let $\sigma_i(t_i) = (m_i^1, m_i^2, m_i^3, m_i^4, m_i^5, m_i^6) \in S_i^{\Gamma(T)}(t_i)$. Suppose, by way of contradiction, that $m_i^6 > 1$. Then, for any profile of messages m_{-i} that agent i's opponents may play, (m_i, m_{-i}) will trigger either Rule 2 or Rule 3. We can partition the message profiles of all agents but i as follows:

$$M_{-i}^2(t_{-i}) \equiv \{m_{-i} \in M_{-i} |\ (m_i, m_{-i}) \text{ triggers Rule 2,}\}$$

and

$$M_{-i}^3(t_{-i}) \equiv \left\{m_{-i} \in M_{-i} \middle|\ (m_i, m_{-i}) \text{ triggers Rule 3.}\right\}$$

Suppose first that type t_i has a belief $\lambda_i \in \Delta(T_{-i} \times M_{-i})$ under which Rule 3 is triggered with positive probability, so that $\sum_{t_{-i}} \sum_{m_{-i} \in M_{-i}^3(t_{-i})} \lambda_i(t_{-i}, m_{-i}) > 0$. If $U_i(m_i^4 | t_i) > U_i(\bar{\alpha} | t_i)$, we define $\hat{m}_i$ as being almost the same as m_i, except that $\hat{m}_i^6$ is chosen to be larger than m_i^6. In doing so, agent i decreases the probability that $\bar{\alpha}$ is chosen in Rule 3. Conditional on Rule 3, this would be a better response, which is a contradiction.

If one has the opposite inequality, note that, under Rule 3, by choosing an appropriate lottery $\hat{m}_i^4$, each agent has a strict incentive to reduce the probability that $\bar{\alpha}$ occurs. This is possible due to Lemma 1. Thus, we can define $\hat{m}_i$ as being almost the same as m_i, except that $\hat{m}_i^4$ is suitably chosen, and $\hat{m}_i^6$ is chosen to be

larger than m_i^6. Similarly, conditional on Rule 3, this would yield a better response, a contradiction.

Therefore, for any rationalizable message profile, the probability that agent i places on Rule 3 occurring is zero.

Now suppose that agent i believes that Rule 2 will be triggered with positive probability, so that $\sum_{t_{-i}} \sum_{m_{-i} \in M_{-i}^2(t_{-i})} \lambda_i(t_{-i}, m_{-i}) > 0$. Thus, there exists exactly one "odd man out." Call him i. Conditional on Rule 2, by NWR, we can suitably choose $\hat{m}_i^3[1], \hat{m}_i^3[2]$, and $\hat{m}_i^6$ large enough, to show that the original message m_i would not be a best response to such beliefs, by constructing a better message $\hat{m}_i$, which is almost the same as m_i, except $\hat{m}_i^3[1], \hat{m}_i^3[2]$, and $\hat{m}_i^6$.

It follows that, in all cases, these choices of $\hat{m}_i$ strictly improve the expected payoff of type t_i if either Rule 2 or Rule 3 is triggered. This implies that m_i is never a best response to any belief λ_i, which contradicts our hypothesis that $m_i \in S_i^{\Gamma(T)}(t_i)$. ■

Step 2: $m_i \in S_i^{\Gamma(T)}(t_i) \Rightarrow \lambda_i^{m_i,t_i}(t_{-i}, m_{-i}) = 0$ for any t_{-i} and any profile (m_i, m_{-i}) under Rules 2 or 3, where $\lambda_i^{m_i,t_i} \in \Delta(T_{-i} \times M_{-i})$ represents the belief held by type t_i to which m_i is a best response.

Proof of Step 2: This follows from Step 1, since every rationalizable message must consist of $m_i^6 = 1$ for every $i \in N$ and the support of the beliefs held by each type must lie in the set of rationalizable profiles. ■

Step 3: $F \subseteq g \circ S^{\Gamma(T)}$. That is, for any $f \in F$, there exists $\sigma_f \in S^{\Gamma(T)}$ such that $g \circ \sigma_f \approx f$.

Proof of Step 3: Let $f \in F$ be arbitrarily chosen, and observe that, by RIC, there exists an extended type profile τ^f such that, for the expanded SCF $\tilde{f}$, $\tilde{f} \circ \hat{t}^{-1}(\tau^f) \approx f(\hat{t}(\tau^f))$. Recall also that, in the direct mechanism associated with $\tilde{f}$, truth-telling is rationalizable for every extended type.

Consider now a message $m_i = (m_i^1, m_i^2, m_i^3, m_i^4, m_i^5, m_i^6)$ of the following type chosen by each type t_i of each agent i in the mechanism Γ:

- for m_i^1, every τ_i such that $\hat{t}_i(\tau_i) = t_i$;
- for m_i^2, every $f \in F$;
- any m_i^3;
- any m_i^4;
- any m_i^5;

- and $m_i^6 = 1$.

We claim that any such message m_i is rationalizable for type t_i. By RIC, if every agent uses any message in this set, the outcomes that result are exactly all the SCFs in F. Also by RIC, and by NWR, every type is at a best response by only using his truthful extended types, knowing that triggering Rule 2 would not result in a better response. By compactness of F, there exists an SCF $f_{t_i}^* \in \arg\max_{f \in F} U_i(f|t_i)$, which maximizes type t_i's interim expected utility over F. Let $\lambda_i^*(t_i)$ denote the belief held by type t_i that the SCF $f_{t_i}^*$, announced by the elected king, say $i+1$ (mod $n+1$), will be the outcome. With such a belief, announcing any SCF in F is a best response for type t_i, since type t_i does not think such SCFs will be chosen in the actual outcome. And the same goes for any other choice of m_i^5, but such announcements will be used to make sure that anyone can be the elected king, thus justifying these beliefs held by every type t_i of every agent i. Finally, while the third and fourth elements of the message are irrelevant under Rule 1, by construction of the mechanism and NWR, type t_i would not have a better reply by inducing Rule 2, hence justifying that $m_i^6 = 1$.

In particular, for the arbitrarily chosen $f \in F$, σ_f can consist of $m^1 = \tau^f$, $m^2 = f$, and any m^5. ■

From now on, we assume that each agent i announces his original type t_i, not the extended type τ_i, in the first component of the message. This is without loss of generality because we only need to extract his extended type τ_i to the extent that it contains his original type t_i and we obtain this simply by taking $\hat{t}_i(\tau_i) = t_i$ where τ_i constitutes the first component of the message. We introduce an additional piece of notation. For any $i \in N$ and any $t_i \in T_i$:

$$\begin{aligned} S_i^{\Gamma(T)}[t_i](t_i) &= \left\{ m_i \in S_i^{\Gamma(T)}(t_i) \,\middle|\, m_i^1 = t_i \text{ and } m_i^6 = 1 \right\} \\ S_i^{\Gamma(T)}[\beta_i](t_i) &= \left\{ m_i \in S_i^{\Gamma(T)}(t_i) \,\middle|\, m_i^1 = \beta_i(t_i) \text{ and } m_i^6 = 1 \right\} \end{aligned}$$

for any deception β.

Now, for each $t \in T$, we define

$$S^{\Gamma(T)}[\beta](t) = \prod_{i \in N} S_i^{\Gamma(T)}[\beta_i](t_i).$$

We say that $\sigma \in S^{\Gamma(T)}[\beta]$ if, for each $t \in T$, $\sigma(t) \in S^{\Gamma(T)}[\beta](t)$. We also define

$$S_i^{\Gamma(T)}(t_i) = S_i^{\Gamma(T)}[t_i](t_i) \cup \left(\bigcup_{\beta \in \mathcal{B}} S_i^{\Gamma(T)}[\beta_i](t_i) \right).$$

And, of course,

$$S^{\Gamma(T)}(t) = \prod_{i \in N} S_i^{\Gamma(T)}(t_i).$$

We denote the identity map by $I_i : T_i \to T_i$ such that $I_i(t_i) = t_i$ for every $t_i \in T_i$.

Step 4: $\forall i \in N : \beta_i \neq I_i \Rightarrow S_i^{\Gamma(T)}[\beta_i] = \emptyset$.

Proof of Step 4: By contradiction, suppose that we have $\bar{\sigma}_j = (\beta_j, \bar{m}_j, \bar{m}_j^3, \bar{m}_j^4, \bar{m}_j^5, 1) \in S_j^{\Gamma(T)}[\beta_j]$ for some $j \in N$ such that $\beta_j(t_j) \neq t_j$ for some $t_j \in T_j$. By Step 2, agent j of every type believes with probability one that Rule 1 is triggered, implying that there exists a deception $\beta \in \mathcal{B}$ such that the set $S_k^{\Gamma(T)}[\beta_k]$ is nonempty for every $k \neq j$. Moreover, for any $k \neq j$, $\bar{\sigma}_k \in S_k^{\Gamma(T)}[\beta_k]$ implies that it is a best response to $\lambda_k \in \Delta(T_{-k} \times M_{-k})$, where the support of this belief consists of strategies that trigger Rule 1 with probability one.

Take the profile $\bar{\sigma} = (\bar{\sigma}_1, \ldots, \bar{\sigma}_n)$ such that $\bar{\sigma}_k \in S_k^{\Gamma(T)}[\beta_k]$ for each $k \in N$. Clearly, by construction, $g \circ \bar{\sigma} \approx \bar{f} \in F \circ \beta$. By RD, we observe that $F \circ \beta \not\subseteq F$.

Therefore, by UBM, there must exist a violation of statement $(**)$, i.e., there must exist $i \in N$, $t_i \in T_i$, $f^* \in F$, and an SCF $y^* : T_{-i} \to \Delta(A)$ such that:

$$(***)\ U_i(f^*|\tilde{t}_i) \geq U_i(y^*|\tilde{t}_i)\ \forall \tilde{t}_i \in T_i \text{ and } U_i(f^* \circ \beta|t_i) < U_i(y^* \circ \beta|t_i).$$

We begin with the following auxiliary claim, whose proof is provided in Section A.1.1 below:

Claim 1 *If $\bar{\sigma} \in S^{\Gamma(T)}[\beta]$, for any $f' \in F \circ \beta$, there exists $\sigma^{f'} \in S^{\Gamma(T)}[\beta]$ such that $g \circ \sigma^{f'} \approx f'$.*

We thus proceed with the proof. By Claim 1, there exists $\sigma^* \in S^{\Gamma(T)}[\beta]$ such that $g \circ \sigma^* \approx f^* \circ \beta$ and $\sigma_k^*(t_k) = (\beta_k(t_k), \sigma_k^{*2}(t_k), \sigma_k^{*3}(t_k), \sigma_k^{*4}(t_k), \sigma_k^{*5}(t_k), 1)$ for each $k \in N$ and $t_k \in T_k$. Since $\sigma_i^*(t_i) \in S_i^{\Gamma(T)}[\beta](t_i)$, there exists $\lambda_i \in \Delta(T_{-i} \times M_{-i})$ such that (i) $\lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$ and (ii) for all $\tilde{m}_i \in M_i$,

$$\sum_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i(g(\sigma_i^*(t_i), m_{-i}); t_i, t_{-i}) \geq \sum_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i(g(\tilde{m}_i, m_{-i}); t_i, t_{-i}).$$

Define

$$\hat{\sigma}_{-i}(\sigma_i^*) \in \arg \max_{(\sigma_i^*, \sigma_{-i}) \in S^{\Gamma(T)}[\beta]} U_i(g \circ (\sigma_i^*, \sigma_{-i})|t_i).$$

The existence of such a degenerate belief $\hat{\sigma}_{-i}(\sigma_i^*)$ is guaranteed by our definition of implementation. (Later, we check this property is indeed satisfied for the mechanism Γ constructed in this proof.) Without loss of generality, we assume that the

winner of the modulo game that yields the SCF $g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*))$ is actually not agent i himself. This can be seen in the proof of Claim 1.

The rest of the proof is intended to establish that $\sigma_i^*(t_i) \notin S_i^{\Gamma(T)}[\beta](t_i)$. By Claim 1, this will imply that $\bar{\sigma}_i(t_i) \notin S_i^{\Gamma(T)}[\beta](t_i)$, which will contradict our initial hypothesis. This will complete the proof of Step 4.

We proceed to detail. Assume that $g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*)) \not\approx f^* \circ \beta$.[15] First, we observe that this assumption implies that $|F \circ \beta| \geq 2$. Since the SCS F satisfies MCI and $|F \circ \beta| \geq 2$, the profile $(\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*))$ is not a Bayesian Nash equilibrium of the game $\Gamma(T)$, since there must exist at least one agent $j \in N \backslash \{i\}$ who has a different strategy that is a better response to the profile $(\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*))$. Since every agent believes with probability one that Rule 1 is triggered, as we have established in Steps 1 to 3, this further implies that there are no strategy profiles in $S^{\Gamma(T)}[\beta]$ that are Bayesian Nash equilibria in the game $\Gamma(T)$. (In particular, σ^* is not a Bayesian Nash equilibrium in the game $\Gamma(T)$ either.)

The preceding argument confirms that $S_{-i}^{\Gamma(T)}[\beta]$ contains multiple strategy profiles, which together with σ_i^*, lead to distinct SCFs. This is consistent with our assumption that $g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*))$ and $f^* \circ \beta$ are two different SCFs, each of which being induced by some rationalizable strategy profile $(\sigma_i^*, \sigma_{-i})$ with $\sigma_{-i} \in S_{-i}^{\Gamma(T)}[\beta]$. We define $\tilde{\lambda}_i \in \Delta(T_{-i} \times M_{-i})$ as follows: $\tilde{\lambda}_i(t_{-i}, m_{-i}) = 0$ if and only if $m_{-i} \neq \hat{\sigma}_{-i}(\sigma_i^*)[t_{-i}]$. We now define $\lambda_i^\varepsilon \in \Delta(T_{-i} \times M_{-i})$ as the belief that assigns probability $1 - \varepsilon$ to $\tilde{\lambda}_i$ and assigns probability ε to the event that all agents other than i use σ_{-i}^*.

By construction, $\sigma_i^*(t_i)$ is a best response to $\tilde{\lambda}_i$. We assume, without loss of generality, that $y^* : T_{-i} \to \Delta(A)$ is the best SCF for type t_i such that $U_i(f^*|\tilde{t}_i) \geq U_i(y^*|\tilde{t}_i)$ for all $\tilde{t}_i$ and $U_i(f^* \circ \beta|t_i) < U_i(y^* \circ \beta|t_i)$.[16] Fix $\varepsilon > 0$ small enough. Since $U_i(g \circ (\sigma_i^*(t_i), \hat{\sigma}_{-i}(\sigma_i^*)) \geq U_i(g \circ (m_i^{'}, \hat{\sigma}_{-i}(\sigma_i^*)|t_i)$ for any $m_i^{'} \in M_i$, the best possible deviation $\sigma_i(t_i) = (m_i^1, m_i^2, m_i^3, m_i^4, m_i^5, m_i^6)$ by agent i of type t_i is to choose $m_i^3[1] = y^*$, $m_i^3[2] = f^*$, and $m_i^6 \to \infty$, keeping the rest of her announcement the same as $\sigma_i^*(t_i)$ so that the SCF is changed only when f^* used to occur under

[15] Later we consider the case where $g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*)) \approx f^* \circ \beta$ and argue that it can also be handled by the same argument we are about to construct.

[16] This is indeed without loss of generality: one could take y^* to be the supremum SCF over the set of SCFs satisfying these inequalities, and then construct the argument below using a sequence of SCFs converging to y^*.

Rule 1. Therefore, for any $m_i^{'} \in M_i$,

$$\begin{aligned}
& \sum_{t_{-i},m_{-i}} \lambda_i^{\varepsilon}(t_{-i},m_{-i}) u_i(g(m_i^{'},m_{-i});t_i,t_{-i}) \\
\leq\ & (1-\varepsilon) U_i(g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon U_i(y^*|t_i) \\
=\ & (1-\varepsilon) U_i(g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon U_i(f^*|t_i) + \varepsilon \left[U_i(y^*|t_i) - U_i(f^*|t_i)\right] \\
=\ & \sum_{t_{-i},m_{-i}} \lambda_i^{\varepsilon}(t_{-i},m_{-i}) u_i(g(\sigma_i^*(t_i),m_{-i});t_i,t_{-i}) + \varepsilon \left[U_i(y^*|t_i) - U_i(f^*|t_i)\right].
\end{aligned}$$

Thus, we obtain that for any $m_i^{'} \in M_i$,

$$\begin{aligned}
& \sum_{t_{-i},m_{-i}} \lambda_i^{\varepsilon}(t_{-i},m_{-i}) u_i(g(\sigma_i^*(t_i),m_{-i});t_i,t_{-i}) \\
\geq\ & \sum_{t_{-i},m_{-i}} \lambda_i^{\varepsilon}(t_{-i},m_{-i}) u_i(g(m_i^{'},m_{-i});t_i,t_{-i}) - \varepsilon \left[U_i(y^*|t_i) - U_i(f^*|t_i)\right].
\end{aligned}$$

Set $\varepsilon^{'} = \varepsilon \left[U_i(y^*|t_i) - U_i(f^*|t_i)\right]$. Hence, if we choose $\varepsilon > 0$ small enough (and hence, $\varepsilon^{'}$ also small enough), we have argued that $\sigma_i^*(t_i)$ is an $\varepsilon^{'}$-best response for type t_i to λ_i^{ε}, even including deviations to messages that trigger Rule 2.

However, we next show the opposite. That is, we now show that $\sigma_i^*(t_i)$ cannot be an $\varepsilon^{'}$-best response for type t_i against λ_i^{ε}. Consider the already described deviation, $\sigma_i(t_i) = (m_i^1, m_i^2, m_i^3, m_i^4, m_i^5, m_i^6)$, by agent i who chooses m_i^6 arbitrarily large, $m_i^3[1] = y^*$, and $m_i^3[2] = f^*$ but keeps the rest of her announcement the same as $\sigma_i^*(t_i)$ so that the SCF is changed only when f^* used to occur under Rule 1. The construction of λ_i^{ε} guarantees that, given $\sigma_i^*(t_i)$, the SCF f^* is realized with probability ε. We define $\{\varepsilon(m_i^6)\}$ as a sequence on $\mathbb{R}$ such that (i) $\varepsilon(m_i^6) > 0$ for each m_i^6; (ii) $\varepsilon(m_i^6) \to 0$ as $m_i^6 \to \infty$; and (iii)

$$\frac{\frac{1}{m_i^6+1}}{\varepsilon(m_i^6)} \to 0 \text{ as } m_i^6 \to \infty$$

For example, we can set $\varepsilon(m_i^6) = 1/\sqrt{m_i^6+1}$, which satisfies the three properties.

Recall that $(\sigma_i^*(t_i), \hat{\sigma}_{-i}(\sigma_i^*))$ induces type t_i's best SCF under any rationalizable strategy profile because type t_i announces 1 in the sixth component of the message (by Step 1) and believes with probability one that other agents also announce 1 in the sixth component of the message (by Step 2). This implies that $U_i(g \circ (\sigma_i^*(t_i), \tilde{\sigma}_{-i}(\sigma_i^*))|t_i) \geq U_i(y^*|t_i)$ because if we have $U_i(g \circ (\sigma_i^*(t_i), \tilde{\sigma}_{-i}(\sigma_i^*))|t_i) < U_i(y^*|t_i)$, type t_i finds it better to induce Rule 2 by announcing a number higher than 1 in the sixth component of the message given the belief that other agents play $\tilde{\sigma}_{-i}(\sigma_i^*)$. This contradicts our Step 1 in which every type announces 1 in the sixth component under rationalizability.

We next show that there exists $\varepsilon > 0$ small enough, such that $\sigma_i^*(t_i)$ is *not* an ε'-best response to λ_i^ε, where $\varepsilon' = \varepsilon\left[U_i(y^*|t_i) - U_i(f^*|t_i)\right]$. Indeed, we confirm this as follows:

$$
\begin{aligned}
& \sum_{t_{-i},m_{-i}} \lambda_i^{\varepsilon(m_i^6)}(t_{-i},m_{-i})u_i(g(\sigma_i(t_i),m_{-i});t_i,t_{-i}) - \varepsilon'(m_i^6) \\
= \;& (1-\varepsilon(m_i^6))\left[\frac{m_i^6}{m_i^6+1}U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i)) + \frac{1}{m_i^6+1}U_i(\bar{y}_i[g\circ(\sigma_i^*,\hat{\sigma}_{-i}(\sigma_i^*))]|t_i)\right] \\
& +\varepsilon(m_i^6)\left[\frac{m_i^6}{m_i^6+1}U_i(y^*;t_i) + \frac{1}{m_i^6+1}U_i(\bar{y}_i[g\circ(\sigma_i^*,\hat{\sigma}_{-i}(\sigma_i^*))]|t_i)\right] - \varepsilon'(m_i^6) \\
& (\because \text{ agent } i \text{ is not the winner of the modulo game under } (\sigma_i,\hat{\sigma}_{-i}(\sigma_i^*)).) \\
\geq \;& \frac{m_i^6}{m_i^6+1}\left[(1-\varepsilon(m_i^6))U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon(m_i^6)U_i(y^*|t_i)\right] \\
& +\frac{1}{m_i^6+1}U_i(\bar{y}_i[g\circ(\sigma_i^*,\hat{\sigma}_{-i}(\sigma_i^*))]|t_i)) - \varepsilon(m_i^6)\left[U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) - U_i(f^*|t_i)\right] \\
& (\because \; \varepsilon'(m_i^6) = \varepsilon(m_i^6)\left[U_i(y^*|t_i) - U_i(f^*|t_i)\right] \text{ and } U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) \geq U_i(y^*|t_i)) \\
\approx \;& (1-\varepsilon(m_i^6))U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon(m_i^6)U_i(y^*|t_i) \\
& -\varepsilon(m_i^6)\left[U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) - U_i(f^*|t_i)\right] \\
& (\text{if we choose } m_i^6 \text{ large enough so that } 1/(m_i^6+1) \to 0 \text{ but } \varepsilon(m_i^6) > 0.) \\
= \;& (1-2\varepsilon(m_i^6))U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon(m_i^6)\left[U_i(y^*|t_i) + U_i(f^*|t_i)\right] \\
= \;& (1-2\varepsilon(m_i^6))U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) + 2\varepsilon(m_i^6)U_i(y^*|t_i) - \varepsilon(m_i^6)\left[U_i(y^*|t_i) - U_i(f^*|t_i)\right] \\
\approx \;& (1-2\varepsilon(m_i^6))U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) + 2\varepsilon(m_i^6)U_i(y^*|t_i) \\
& (\text{if we choose } \varepsilon(m_i^6) \text{ small enough, noting } 2U_i(y^*|t_i) > U_i(y^*|t_i) - U_i(f^*|t_i)) \\
\approx \;& (1-\varepsilon(m_i^6))U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon(m_i^6)U_i(y^*|t_i) \\
& (\text{if we choose } m_i^6 \text{ large enough so that } \varepsilon(m_i^6) \to 0) \\
> \;& (1-\varepsilon(m_i^6))U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon(m_i^6)U_i(f^*|t_i) \\
& (\because \; U_i(y^*|t_i) > U_i(f^*|t_i), \text{ and } \varepsilon(m_i^6) > 0.) \\
= \;& \sum_{t_{-i},m_{-i}} \lambda_i^{\varepsilon(m_i^6)}(t_{-i},m_{-i})u_i(g(\sigma_i^*(t_i),m_{-i});t_i,t_{-i}).
\end{aligned}
$$

Hence, we have established the desired opposite inequality, showing that $\sigma_i^*(t_i)$ is not an ε'-best response for type t_i against λ_i^ε.[17]

[17]To make our argument more transparent, we could divide it into the following two cases. We first assume $U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) > U_i(y^*|t_i)$. In this case, we immediately obtain a strict inequality even before we take $1/(m_i^6+1) \to 0$. Otherwise, i.e., if $U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) = U_i(y^*|t_i)$, when we obtain $(1-2\varepsilon(m_i^6))U_i(g\circ(\sigma_i^*(t_i),\hat{\sigma}_{-i}(\sigma_i^*))|t_i) + 2\varepsilon(m_i^6)U_i(y^*|t_i)$, this is

For the case where $g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*)) \approx f^*$, we first observe that $\sigma_i^*(t_i)$ is a best response to λ_i^ε independently of the size of ε because $g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*)) \approx g \circ \sigma^* \approx f^*$. We next claim that if we choose m_i^6 large enough, the same deviation strategy $\sigma_i(t_i)$ constructed above is a better response to λ_i^ε than $\sigma_i^*(t_i)$. Specifically, given the belief λ_i^ε, $\sigma_i(t_i)$ induces the SCF y^* with probability $m_i^6/(m_i^6+1)$ and the SCF $\bar{y}_i[f^*]$ with the rest of probability. Since $U_i(y^*|t_i) > U_i(f^*|t_i)$, by choosing m_i^6 large enough, we obtain

$$\sum_{t_{-i}, m_{-i}} \lambda_i^\varepsilon(t_{-i}, m_{-i}) u_i(g(\sigma_i(t_i), m_{-i}); t_i, t_{-i}) > \sum_{t_{-i}, m_{-i}} \lambda_i^\varepsilon(t_{-i}, m_{-i}) u_i(g(\sigma_i^*(t_i), m_{-i}), t_i, t_{-i}).$$

Thus, even if $g \circ (\sigma_i^*, \hat{\sigma}_{-i}(\sigma_i^*)) \approx f^*$, we obtain the desired contradiction, as in the previous case. Hence, regardless of whether $g \circ (\sigma_i^*, \hat{\sigma}_{-i}) \not\approx f^*$ or $g \circ (\sigma_i^*, \hat{\sigma}_{-i}) \approx f^*$, we conclude that $\sigma_i^* \notin S_i^{\Gamma(T)}[\beta]$. This completes the proof of Step 4. ■

Now, we shall complete the proof of Theorem 4. By Step 4, it follows that for any $t \in T^*$,

$$S^{\Gamma(T)}(t) = \prod_{i \in N} S_i^{\Gamma(T)}(t_i) = \prod_{i \in N} \left(S_i^{\Gamma(T)}[t_i](t_i) \cup \left(\bigcup_{\beta \in \mathcal{B}} S_i^{\Gamma(T)}[\beta_i](t_i) \right) \right) = \prod_{i \in N} S_i^{\Gamma(T)}[t_i](t_i).$$

This, together with Step 3, further implies that for any $t \in T^*$,

$$\bigcup_{m \in S^{\Gamma(T)}(t)} \{g(m)\} = F(t).$$

Finally, we need to show that in this canonical mechanism, for any $i \in N$, $t_i \in T_i$, and $\sigma_i \in S_i^{\Gamma(T)}$, there exist a belief $\lambda_i^{\sigma_i(t_i)} \in \Delta(T_{-i} \times M_{-i})$ and profile of pure strategies $\sigma_{-i} \in S_{-i}^{\Gamma(T)}$ such that $\lambda_i^{\sigma_i(t_i)}(t_{-i}, \sigma_{-i}(t_{-i})) = 1$ for each $t_{-i} \in T_{-i}$ and $\sigma_i(t_i)$ is a best response against $\lambda_i^{\sigma_i(t_i)}$. Throughout Steps 1 through 4, every type believes with probability one that Rule 1 is triggered (i.e., $m_i^6 = 1$ for each $i \in N$) and all agents truthfully announce their type in the first component of the message. This implies that m_i^3 and m_i^4 do not matter under rationalizability. Therefore, each type t_i's belief concentrated over $T_{-i} \times M_i^3 \times M_i^4$ can be made degenerate without loss of generality. For each type t_i, we define $f_i^* \in \arg\max_{\tilde{f} \in F} U_i(\tilde{f}|t_i)$. Then, type t_i's belief, λ_i can be made degenerate over $T_{-i} \times M_{-i}^2 \times M_{-i}^5$ such that some other player j becomes the winner of the modulo game and the winner of the modulo game chooses $m_j^2 = f_i^*$ with probability one. Hence, the desired property (2) of implementability is satisfied for the mechanism we have constructed.

This concludes the proof of Theorem 4. ■

equivalent to $(1 - \varepsilon(m_i^6))U_i(g \circ (\sigma_i^*(t_i), \hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon(m_i^6)U_i(y^*|t_i)$. This is larger than $(1 - \varepsilon(m_i^6))U_i(g \circ (\sigma_i^*(t_i), \hat{\sigma}_{-i}(\sigma_i^*))|t_i) + \varepsilon(m_i^6)U_i(f^*|t_i)$, regardless of the size of $\varepsilon(m_i^6)$.

A.1.1. Proof of Claim 1

We set $n+1 \equiv 1$ and $0 \equiv n$. Fix $f^{'} \in F \circ \beta$. We minimally modify $\bar{\sigma}$ to construct a strategy profile $\sigma^{f^{'}}$, which induces Rule 1 with probability one, and in which all agents announce their type via β in the first component of their message, and agent $i+1$ becomes the winner of the modulo game, as follows. First, for agent $i+1$, define $\sigma_{i+1}^{f^{'}}(t_{i+1}) = (\beta_{i+1}(t_{i+1}), \sigma_{i+1}^{f^{'},2}(t_{i+1}), \bar{\sigma}_{i+1}^{3}(t_{i+1}), \bar{\sigma}_{i+1}^{4}(t_{i+1}), i+1, 1)$ for each type $t_{i+1} \in T_i$, where $\sigma_{i+1}^{f^{'},2}(t_{i+1}) = f^{'}$ for each t_{i+1}. Second, for each $j \in N \backslash \{i+1\}$, define $\sigma_j^{f^{'}}(t_j) = (\beta_j(t_j), \sigma_j^{f^{'},2}(t_j), \bar{\sigma}_j^{3}(t_j), \bar{\sigma}_j^{4}(t_j), 1, 1)$ for each t_j such that we remain ambiguous about the choice of $\sigma_j^{f^{'},2}(t_j) \in F \circ \beta$. Even if we are largely ambiguous about the specification of the second component of the message for each agent $j \neq i+1$, our construction of $\sigma^{f^{'}}$ guarantees that $g \circ \sigma^{f^{'}} \approx f^{'}$. Agent $i+1$ of each type t_{i+1} believes with probability one that someone else will be the winner of the modulo game and the winner of the modulo game will choose the best one among all SCFs in $F \circ \beta$ for type t_{i+1}. Moreover, each agent $j \neq i+1$ of each type t_j believes with probability one that someone else will be the winner of the modulo game, and that the winner of the modulo game will choose the best one among all SCFs in $F \circ \beta$ for type t_j. What remains to show is that $\sigma^{f^{'}} \in S^{\Gamma(T)}[\beta]$. We proceed to do so.

First, we show that $\sigma_{i+1}^{f^{'}}(t_{i+1})$ is a best response for type t_{i+1} to some belief. Define $\lambda_{i+1}^{*} \in \Delta(T_{-(i+1)} \times M_{-(i+1)})$ with support of the following property: type t_{i+1} believes with probability one that someone else will be the winner of the modulo game, and that the winner of the modulo game will choose the best one f_{i+1} among all SCFs for type t_{i+1}, i.e.,

$$f_{i+1} \in \arg \max_{\tilde{\beta}_{-(i+1)}:(\beta_{i+1},\tilde{\beta}_{-(i+1)}) \in \mathcal{B}, \tilde{f} \in F \circ (\beta_{i+1},\tilde{\beta}_{-(i+1)})} U_{i+1}(\tilde{f}|t_{i+1}).$$

Assume $j \in N \backslash \{i+1\}$. We show that $\sigma_j^{f^{'}}(t_j)$ is a best response for type t_j to some belief. Define $\lambda_j^{*} \in \Delta(T_{-j} \times M_{-j})$ with support of the following property: each agent $j \neq i+1$ of each type t_j believes with probability one that someone else will be the winner of the modulo game, and that the winner of the modulo game will choose the best one f_j among all SCFs for type t_j, i.e.,

$$f_j \in \arg \max_{\tilde{\beta}_{-j}:(\beta_j,\tilde{\beta}_{-j}) \in \mathcal{B}, \tilde{f} \in F \circ (\beta_j,\tilde{\beta}_{-j})} U_j(\tilde{f}|t_j).$$

In sum, we conclude that we have $\sigma^{f^{'}} \in S^{\Gamma(T)}[\beta]$ such that $g \circ \sigma^{f^{'}} \approx f^{'}$, as desired. ■

A.2. Proof of Theorem 5

Necessity of RIC and closure follows from Theorems 1 and 2, respectively. We are left with sufficiency. Sufficiency follows from a variant of the canonical mechanism constructed in the proof of Theorem 4, as we explain next.

First, we show the following proposition, establishing that, in these economic environments, an important preference-nestedness condition involving deceptions across types is impossible.

Proposition 4 *Suppose that an SCS F contains a continuous SCF $f \in F$ such that for every agent $i = 1, 2$, (i) $U_i(f|0) = 0$, and (ii) for $t_i \in (0, 1]$, $U_i(f|t_i) > 0$. Then, for any deception $\beta \in \mathcal{B}$, the following preference-nestedness condition can never hold:*

$$(PN) \left[\begin{array}{l} U_i(f|\tilde{t}_i) \geq U_i(y|\tilde{t}_i) \ \forall \tilde{t}_i \in T_i \\ \Rightarrow U_i(f \circ \beta|t_i) \geq U_i(y \circ \beta|t_i) \end{array} \right], \forall i \in N, \ \forall t_i \in T_i.$$

Proof of Proposition 4: Throughout the proof of the proposition, we focus on a continuous SCF $f \in F$ satisfying assumptions (i) and (ii) above. First, we have a profile of bidding strategies underlying f, denoted by $\sigma = (\sigma_1, \sigma_2)$. Fixing $i \in \{1, 2\}$ arbitrarily, $\sigma_i : [0, 1] \to \mathbb{R}$ is agent i's continuous bidding function, and denoting by g the outcome function corresponding to the first-price auction, we can write that $g \circ \sigma \approx f$. Next, we identify the SCFs $y : T_{-i} \to \Delta(A)$ for which the first clause of (PN) holds, i.e.,

$$U_i(f|\tilde{t}_i) \geq U_i(y|\tilde{t}_i) \ \forall \tilde{t}_i \in T_i.$$

Since this inequality is imposed on all types, it is also imposed on $\tilde{t}_i = 0$. Since, by assumption (i) in the proposition, $U_i(f|0) = 0$, the expected utility of y must also be at most zero. Since y does not depend on agent i's type, it must be that y consists of outcomes where the good is always allocated to agent $j \neq i$, or if it is allocated to agent i, it must be sold to i for a price of 1 (in this way, no type can ever get positive expected utility under the SCF y).

The rest of the proof of the proposition is completed by two claims.

Claim 2 *For any deception $\beta \in \mathcal{B}$, if $\beta_i(0) \neq 0$ for some $i = 1, 2$, Condition (PN) cannot hold.*

Proof of Claim 2: Choose a deception β such that $\beta_i(0) \neq 0$ for some $i = 1, 2$. We set $\beta_i(0) = \bar{t}_i \in (0, 1]$. Consider now the second clause in (PN) for type $t_i = 0$. For (PN) to hold, we must have:

$$U_i(f \circ \beta|0) \geq U_i(y \circ \beta|0).$$

However, although $U_i(y \circ \beta|0) = 0$, we now show that $U_i(f \circ \beta|0) < 0$.

By assumption (ii) in the statement of the proposition, type $\bar{t}_i$'s expected utility from f is:

$$U_i(f|\bar{t}_i) = (\bar{t}_i - \sigma_i(\bar{t}_i)) \int_{[0,1]} \text{Prob}\{\sigma_i(\bar{t}_i) \geq \sigma_j(t_j)\} dt_j > 0.$$

This implies that the probability that this type gets the good is positive, i.e., $\int_{[0,1]} \text{Prob}\{\sigma_i(\bar{t}_i) \geq \sigma_j(t_j)\} dt_j > 0$, further implying that $\sigma_i(\bar{t}_i) > 0$.

With the deception β, type $t_i = 0$'s expected utility from $f \circ \beta$ is precisely:

$$U_i(f \circ \beta|0) \;\; = \;\; -\sigma_i(\bar{t}_i) \int_{[0,1]} \text{Prob}\{\sigma_i(\bar{t}_i) \geq \sigma_j(t_j)\} dt_j < 0.$$

We thus obtain the desired inequality. ■

Claim 3 *For any deception $\beta \in \mathcal{B}$, if $\beta_i(0) = 0$ for both $i = 1, 2$, there exists $j \in \{1, 2\}$ and type $t_j \in [0, 1]$ such that $U_j(f \circ \beta|t_j) < 0$. Thus, condition (PN) cannot hold for β.*

Proof of Claim 3: Suppose, by way of contradiction, that there exists a deception $\beta \in \mathcal{B}$ such that $\beta_i(0) = 0$ and $U_i(f \circ \beta|t_i) \geq 0$ for every $i = 1, 2$ and $t_i \in [0, 1]$. Since β is a deception, there must exist $i \in N$ and type $t_i \in (0, 1]$ such that $\beta_i(t_i) \neq t_i$. We focus on such agent i in the rest of the argument. We define

$$t_i[\beta] = \inf\{\tilde{t}_i \in [0, 1] : \beta_i(\tilde{t}_i) \neq \tilde{t}_i\}$$

We then establish the following key auxiliary lemma:

Lemma 2 *Suppose that there exists a deception $\beta \in \mathcal{B}$ such that $\beta_i(0) = 0$ and $U_i(f \circ \beta|t_i) \geq 0$ for every $i = 1, 2$ and $t_i \in [0, 1]$. Then, there exists a deception $\beta' \in \mathcal{B}$, $\beta'_j = \beta_j$ for $j \neq i$, such that $t_i[\beta'] = t_i[\beta]/2$, defined as follows: there exists $\varepsilon > 0$ small enough so that $\beta'_i(t_i[\beta]/2) = t_i[\beta]/2 - \varepsilon > 0$; $\beta'_i(t_i) = \beta_i(t_i)$ for any other type t_i; and $U_i(f \circ \beta'|t_i) \geq 0$ for every $t_i \in [0, 1]$.*

Proof of Lemma 2: By construction of β', we have $U_i(f \circ \beta'|t_i) \geq 0$ for every $t_i \neq t_i[\beta]/2$. By definition of $t_i[\beta]$, we have $\beta_i(t_i[\beta]/2) = t_i[\beta]/2$. By assumption (ii) in the statement of the proposition, we have that $U_i(f \circ \beta|t_i[\beta]/2) > 0$. By the continuity of f and of expected utility, we can choose $\varepsilon > 0$ small enough so that $U_i(f \circ \beta'|t_i[\beta]/2) > 0$. Since β' is the same as β except for type $t_i[\beta'] = t_i[\beta]/2$, in particular agent $j \neq i$ still wins with the same probability, paying the same price. We thus have that $U_i(f \circ \beta'|t_i) \geq 0$ for all $t_i \in [0, 1]$, as desired. ■

With the repeated application of this lemma, one builds a loop that eventually leads to a contradiction. Specifically, one can construct a pair of sequences, $\{\beta^k\}_{k=0}^{\infty}$ and $\{t_i[\beta^k]\}_{k=0}^{\infty}$, such that $\beta^0 = \beta$, satisfying $\beta_i^k(0) = 0$ and $U_i(f \circ \beta^k | t_i) \geq 0$ for every $i = 1, 2$ and $t_i \in [0, 1]$, and $t_i[\beta^{k+1}] = (1/2)^k t_i[\beta]$ for each nonnegative integer $k \geq 0$. By construction, if one chooses K large enough, we have that $t_i[\beta^K]$ is arbitrarily close to zero.

Consider now the second clause in (PN) for type $t_i = t_i[\beta^K] - \varepsilon > 0$. For (PN) to hold, we must have:

$$U_i(f \circ \beta^{K+1} | t_i[\beta^K] - \varepsilon) \geq U_i(y \circ \beta^{K+1} | t_i[\beta^K] - \varepsilon).$$

We know that $U_i(y \circ \beta^{K+1} | t_i[\beta^K] - \varepsilon) = 0$, but we now show that $U_i(f \circ \beta^{K+1} | t_i[\beta^K] - \varepsilon) < 0$.

Once again, by assumption (ii) in the statement of the proposition, type $t_i[\beta^K] - \varepsilon$ receives positive expected utility from f:

$$U_i(f | t_i[\beta^K] - \varepsilon) = (t_i[\beta^K] - \varepsilon - \sigma_i(t_i[\beta^K] - \varepsilon)) \int_{[0,1]} \text{Prob}\{\sigma_i(t_i[\beta^K] - \varepsilon) \geq \sigma_j(t_j)\} dt_j > 0.$$

This implies that $\int_{[0,1]} \text{Prob}\{\sigma_i(t_i[\beta^K] - \varepsilon) \geq \sigma_j(t_j)\} dt_j > 0$, which in turn implies that $\sigma_i(t_i[\beta^K] - \varepsilon) > 0$. On the other hand, recalling that type $t_i[\beta^K]$ is arbitrarily close to 0, this type's expected utility from $f \circ \beta^{K+1}$ is:

$$U_i(f \circ \beta^{K+1} | t_i[\beta^K]) \;\; = \;\; -\sigma_i(t_i[\beta^K] - \varepsilon) \int_{[0,1]} \text{Prob}\{\sigma_i(t_i[\beta^K] - \varepsilon) \geq \sigma_j(t_j)\} dt_j < 0.$$

We thus obtain the desired inequality, a violation of condition (PN) for deception β^{K+1}. Since the lemma above can be read as follows: "if there exists a deception β^k with some properties, there exists a deception β^{k+1} with the same properties," we can unravel its conclusions, starting from β^{K+1} and going backwards in the sequence, to conclude that $\beta^0 = \beta$ does not have the assumed properties. We therefore have that, either $\beta_i(0) \neq 0$, or there is a type $t_i \in [0, 1]$ such that $U_i(f \circ \beta | t_i) < 0$. The former is impossible, as it would contradict Claim 2, so we must have the latter, and hence, the proof of Claim 3 is complete. ■

By Claims 2 and 3, we conclude that (PN) can never hold for any deception β. This completes the proof of Proposition 4. ■

To continue with the proof of Theorem 5, we modify slightly the proof of the general sufficiency theorem in section 6, i.e., Theorem 4.[18] We note that the MCI

[18]Or more precisely, its extension in the appendix (Section A.3), which takes care of environments where type spaces are compact subsets of Euclidean spaces, such as intervals in the real line, as is the case here.

condition is trivially satisfied in these environments, as, after the assumed f is in the SCS, there is no SCF that simultaneously can be the maximizer of every type and every agent's expected utility within F, given the definition of the set of alternatives in the first-price auction. Furthermore, the NWR condition, used to impose bad outcomes after deviations, can be dispensed with, by always using "Zero" as the bad outcome in Rule 2 of the implementing mechanism, i.e., the good is allocated to the deviating agent with probability zero. Similarly, outcome $\bar{\alpha}$ in Rule 3 can be replaced with a collection of outcomes that use the "Zero" outcome as well as others where the good is assigned to each agent with small uniform probability.

After these changes are made, note that Steps 1, 2, and 3 in the proof of Theorem 4 go through with no change. The only modification in the proof of Step 4 uses Proposition 4, instead of the RD and UBM conditions, to show the existence of the preference reversal $(***)$, for any deception β. No further change is needed, and the proof of Theorem 5 is complete. ■

A.3. Extension to a More General Setup

We extend this paper's analysis to a more general environment with incomplete information. Assume that T_i is a *Polish* space T_i associated with its Borel σ-algebra $\mathscr{T}_i$. We endow T_{-i} and T with the product Borel sigma-algebras $\mathscr{T}_{-i}$ and $\mathscr{T}$, respectively. Note that T_{-i} and T are also Polish spaces. Let $\Delta(T_{-i})$ denote the set of probability distributions on the measurable space $(T_{-i}, \mathscr{T}_{-i})$ endowed with the weak* topology. Each agent i's system of "interim" beliefs is expressed as a $\mathscr{T}_i$-measurable function $\pi_i : T_i \to \Delta(T_{-i})$. Then, we call $(T_i, \mathscr{T}_i, \pi_i)_{i\in N}$ a *type space*. Let A denote the set of pure outcomes associated with its sigma-algebra $\mathscr{A}$ containing all singleton sets. Let $\Delta(A)$ be the set of probability distributions over measurable space $(A, \mathscr{A})$. Agent i's state dependent von Neumann-Morgenstern utility function is denoted $u_i : \Delta(A)\times T \to \mathbb{R}$, which is assumed to be a $\mathscr{A}\times\mathscr{T}$-measurable function. We can now define an *environment* as $\mathcal{E} = (A, \mathscr{A}, \{u_i, T_i, \mathscr{T}_i, \pi_i\}_{i\in N})$.

A subset of T is called an event if it is $\mathscr{T}$-measurable.[19] An event $E = E_1 \times \cdots \times E_n \subseteq T$ is said to be *belief-closed* if, for each $i \in N$ and $t_i \in E_i$, we have $\pi_i[t_i](E_{-i}) = 1$. We assume that the planner only cares about the belief-closed subset of the type space $(T_i^*, \mathscr{T}_i^*)_{i\in N}$ where $T_i^* \subseteq T_i$ and $\mathscr{T}_i^*$ is its relative sigma-algebra for every $i \in N$.

A (stochastic) *social choice function* (SCF) is a $\mathscr{T}$-measurable function $f : T \to \Delta(A)$. Let $\mathbb{F}$ be the collection of all $\mathscr{T}$-measurable SCFs. A *social choice set* (SCS) F is defined as a nonempty compact subset of $\mathbb{F}$. Two SCSs F and H

[19] Since $\mathscr{T}$ is the product measure, any event constitutes a product set.

are said to be *equivalent* ($F \approx H$) if there exists a bijection $\xi : F \to H$ such that $\sup\{|f(\mathcal{A}|t) - h(\mathcal{A}|t)| : t \in T^*, \mathcal{A} \in \mathscr{A}\} = 0$ for every $f \in F$ and every $h \in H$ satisfying $h = \xi(f)$. This means that the two SCSs "coincide" for every $t \in T^*$.

A *mechanism* (or *game form*) $\Gamma = ((M_i, \mathscr{M}_i)_{i\in N}, g)$ describes a nonempty message space M_i for each agent i, equipped with a sigma-algebra $\mathscr{M}_i$ and an $\mathscr{M}$-measurable outcome function $g : M \to \Delta(A)$, where $M = \times_{i\in N} M_i$ is associated with product sigma-algebra $\mathscr{M}$.

The interim expected utility of agent i of type t_i that pretends to be of type t_i' in the direct-revelation mechanism associated with an SCF f, provided all other agents are truthful, is defined as:

$$U_i(f; t_i'|t_i) \equiv \int_{\mathscr{T}_{-i}} u_i(f(t_i', t_{-i}); (t_i, t_{-i}))\pi_i[t_i](dt_{-i}).$$

Let $U_i(f|t_i) = U_i(f; t_i|t_i)$.

A.3.1. Implementation in Rationalizable Strategies

Given a mechanism $\Gamma = (M, \mathscr{M}, g)$, let $\Gamma(T)$ denote an incomplete information game associated with a type space $(T_i, \mathscr{T}_i, \pi_i)_{i\in N}$. Let $\sigma_{-i} : T_{-i} \to \Delta(M_{-i})$ denote a $\mathscr{T}_{-i}$-measurable randomized strategy profile of all agents other than i and Σ_{-i} the set of randomized strategies, where $\Delta(M_{-i})$ denotes the set of probability measures over $(M_{-i}, \mathscr{M}_{-i})$ endowed with the weak* topology. We assume that $g \circ (m_i, \sigma_{-i})$ is a $\mathscr{T} \times \mathscr{M}$-measurable function and $g \circ (m_i, \sigma_{-i}) \in \mathbb{F}$ for every $m_i \in M_i$ and $\sigma_{-i} \in \Sigma_{-i}$. With abuse of notation, we let

$$U_i(g \circ (m_i, \sigma_{-i})|t_i) \equiv \int_{\mathscr{T}_{-i}\times\mathscr{M}_{-i}} u_i(g(m_i, m_{-i}); (t_i, t_{-i}))\sigma_{-i}(dm_{-i}|t_{-i})\pi_i[t_i](dt_{-i}).$$

We fix a mechanism $\Gamma = (M, \mathscr{M}, g)$ and define a message correspondence profile $S = (S_1, \ldots, S_n)$, where each $S_i : T_i \to 2^{M_i}$, which is $\mathscr{M}_i$-measurable and we write $\mathcal{S}$ for the collection of message correspondence profiles. The collection $\mathcal{S}$ is a lattice with the natural ordering of set inclusion: $S \leq S'$ if $S_i(t_i) \subseteq S_i'(t_i)$ for all $i \in N$ and $t_i \in T_i$. The largest element is $\bar{S} = (\bar{S}_1, \ldots, \bar{S}_n)$, where $\bar{S}_i(t_i) = M_i$ for each $i \in N$ and $t_i \in T_i$. The smallest element is $\underline{S} = (\underline{S}_1, \ldots, \underline{S}_n)$, where $\underline{S}_i(t_i) = \emptyset$ for each $i \in N$ and $t_i \in T_i$.

We define an operator b to iteratively eliminate never best responses. The operator $b : \mathcal{S} \to \mathcal{S}$ is now defined as: for every $i \in N$ and $t_i \in T_i$,

$$b_i(S)[t_i] \equiv \left\{ m_i \;\middle|\; \begin{array}{l} \exists \mathscr{T}_{-i} \times \mathscr{M}_{-i}\text{-measurable } \lambda_i \in \Delta(T_{-i} \times M_{-i}) \text{ such that} \\ (1)(t_{-i}, m_{-i}) \in \text{supp}(\lambda_i) \Rightarrow m_{-i} \in S_{-i}(t_{-i}); \\ (2)\int_{\mathscr{M}_{-i}} \lambda_i(t_{-i}, dm_{-i}) = \pi_i(t_i)[t_{-i}]; \\ (3) m_i \in \arg\max_{m_i'} \int_{\mathscr{T}_{-i}\times\mathscr{M}_{-i}} u_i(g(m_i', m_{-i}); t_i, t_{-i})\lambda_i(dt_{-i}, dm_{-i}) \end{array} \right\}.$$

Observe that b is increasing by definition: i.e., $S \leq S' \Rightarrow b(S) \leq b(S')$. By Tarski's fixed point theorem, there is a largest point of b, which we label $S^{\Gamma(T)}$. Thus, (i) $b(S^{\Gamma(T)}) = S^{\Gamma(T)}$ and (ii) $b(S) = S \Rightarrow S \leq S^{\Gamma(T)}$. We can also construct the fixed point $S^{\Gamma(T)}$ by starting with $\bar{S}$ – the largest element of the lattice – and iteratively applying the operator b.

Because the mechanism Γ is infinite, transfinite induction may be necessary to reach the fixed point. It is useful to define

$$S_i^{\Gamma(T),k}(t_i) \equiv b_i\left(b^{k-1}\left(\bar{S}\right)\right)[t_i],$$

using transfinite induction if necessary. Thus $S_i^{\Gamma(T)}(t_i)$ are the sets of messages surviving (transfinite) iterated deletion of never best responses of type t_i of agent i. We denote by σ_i a $\mathscr{T}_i$-measurable selection from $S_i^{\Gamma(T)}$ and call it a rationalizable strategy of player i. We acknowledge the following structure of $S^{\Gamma(T)}$:

$$S^{\Gamma(T)} = \prod_{i \in N} S_i^{\Gamma(T)}.$$

Next, we provide the definitions of *rationalizable implementation* that we use in the paper.

Definition 14 (Full Implementation in Rationalizable Strategies) *An SCS F is **fully implementable in rationalizable strategies** if there exists a mechanism $\Gamma = (M, \mathscr{M}, g)$ with the following three properties: (i) for each $t \in T^*$,*

$$\bigcup_{m \in S^{\Gamma(T)}(t)} \{g(m)\} = F(t),$$

(ii) for any $i \in N$, $t_i \in T_i$, and $\sigma_i \in S_i^{\Gamma(T)}$, there exist a $\mathscr{T}_{-i} \times \mathscr{M}_{-i}$-measurable belief $\lambda_i^{\sigma_i(t_i)} \in \Delta(T_{-i} \times M_{-i})$ and $\mathscr{T}_{-i}$-measurable $\sigma_{-i} \in S_{-i}^{\Gamma(T)}$ such that $\lambda_i^{\sigma_i(t_i)}(t_{-i}, \sigma_{-i}(t_{-i})) = 1$ a.s. for each $t_{-i} \in T_{-i}$ and $\sigma_i(t_i)$ is a best response against $\lambda_i^{\sigma_i(t_i)}$, and (iii) for any $\sigma', \sigma'' \in S^{\Gamma(T)}$ and $T', T'' \subseteq T$ for which $T' \cup T'' = T$ and $T' \cap T'' = \emptyset$, there exists $f \in F$ such that for any $t \in T^*$,

$$f(t) = \begin{cases} g(\sigma'(t)) & \text{if } t \in T' \\ g(\sigma''(t)) & \text{if } t \in T'' \end{cases}$$

Definition 15 (Weak Implementation in Rationalizable Strategies) *An SCS F is **weakly implementable in rationalizable strategies** if there exists a mechanism $\Gamma = (M, \mathscr{M}, g)$ with the following three properties: (i) for each $t \in T^*$,*

$$\emptyset \neq \bigcup_{m \in S^{\Gamma(T)}(t)} \{g(m)\} \subseteq F(t),$$

and conditions (ii) and (iii) as above.

A.3.2. Necessity for Rationalizable Implementation

This subsection discusses three necessary conditions: (1) rationalizable incentive compatibility (RIC); (2) closure; and (3) uniform Bayesian monotonicity (UBM). The proofs of the corresponding results are exactly as provided in the main text. When dealing with a more general setup, no modification for RIC is needed. The only modification we need for defining closure in a more general case is the measurability requirement for events. More specifically, a subset of T is said to be an event if it is $\mathscr{T}$-measurable. Finally, the only modification one needs for UBM is the requirement that each deception $\beta_i : T_i \to T_i$ is $\mathscr{T}_i$-measurable.

A.3.3. Sufficiency for Rationalizable Implementation

In this section, we discuss how one can extend Theorem 4 to a more general setup. For each SCF $f \in \mathbb{F}$, define

$$Y_i[f] \equiv \left\{ y_i : T_{-i} \to \Delta(A) \;\middle|\; \begin{array}{l} y_i \text{ is } \mathscr{T}_{-i}\text{-measurable and} \\ U_i(f|\tilde{t}_i) \geq U_i(y_i|\tilde{t}_i) \;\forall \tilde{t}_i \in T_i^* \end{array} \right\}.$$

The set $Y_i[f]$ is associated with its Borel σ-algebra $\mathscr{Y}_i[f]$.

Since T_i is a Polish space, $\Delta(T_i)$ can also be made Polish. We denote by $\{t_i^\ell\}_{\ell=1}^\infty$ its countable dense subset of $\Delta(T_i)$. Similarly, since $T_{-i} \times T_{-i}$ is a Polish space, $\Delta(T_{-i} \times T_{-i})$ can also be made Polish. So, we denote by $\{\psi_i^k\}_{k=1}^\infty$ its countable dense subset of $\Delta(T_{-i} \times T_{-i})$. Since F satisfies NWR, for each $f \in F$ and $i \in N$, we define the uniform SCF $\bar{y}_i[f]$ as follows: there exist $\delta, \eta \in (0,1)$ such that

$$\bar{y}_i[f] \equiv \frac{(1-\delta)(1-\eta)}{2} \sum_{\ell=1}^\infty \eta^{\ell-1} \sum_{k=1}^\infty \delta^{k-1} \left\{ y_i'[f; t_i^\ell, \psi_i^k] + y_i[f; t_i^\ell, \psi_i^k] \right\}.$$

Recall that this uniform SCF $\bar{y}_i[f]$ is used in the canonical mechanism for Theorem 4. Note also that $\{t_i^\ell\}_{\ell=1}^\infty$ is a dense subset of $\Delta(T_i)$ and $\{\psi_i^k\}_{k=1}^\infty$ is a dense subset of $\Delta(T_{-i} \times T_{-i})$, respectively. As expected utility is continuous in both $\Delta(T_i)$ and $\Delta(T_{-i} \times T_{-i})$, NWR together with the uniform SCF $\bar{y}_i[f]$ plays exactly the same role in the proof of Theorem 4 as if the type space is countable.

Similarly, for each $i \in N$, we define the uniform lottery $\bar{\alpha}_i \in \Delta(A)$ as follows: there exist $\delta, \eta \in (0,1)$ such that

$$\bar{\alpha}_i \equiv \frac{(1-\delta)(1-\eta)}{2} \sum_{\ell=1}^\infty \eta^{\ell-1} \sum_{k=1}^\infty \delta^{k-1} \left\{ \alpha_i'[t_i^\ell, \phi_i^k] + \alpha_i[t_i^\ell, \phi_i^k] \right\}.$$

Finally, we define

$$\bar{\alpha} \equiv \frac{1}{n} \sum_{i \in N} \bar{\alpha}_i.$$

Recall that this uniform lottery $\bar{\alpha}$ is used in the canonical mechanism of Theorem 4. Note also that $\{t_i^\ell\}_{\ell=1}^\infty$ is a dense subset of $\Delta(T_i)$ and $\{\psi_i^k\}_{k=1}^\infty$ is a dense subset of $\Delta(T_{-i} \times T_{-i})$, respectively. Once again, as expected utility is continuous in both $\Delta(T_i)$ and $\Delta(T_{-i} \times T_{-i})$, NWR (more precisely, Lemma 1) together with the uniform lottery $\bar{\alpha}$ plays exactly the same role in the proof of Theorem 4 as if the type space is countable.

We thus state the extension of Theorem 4 to a more general setup:

Theorem 6 *If an SCS F satisfies rationalizable incentive compatibility, uniform Bayesian monotonicity, closure, NWR, MCI, and RD, then it is fully implementable in rationalizable strategies.*

Proof: We use the same mechanism proposed in the proof of Theorem 4. In the proposed mechanism $\Gamma = (M, \mathscr{M}, g)$, each agent i sends a message $m_i = (m_i^1, m_i^2, m_i^3, m_i^4, m_i^5, m_i^6) \in \prod_{k=1}^6 M_i^k = M_i$ where $m_i^1 \in \mathcal{T}_i$, $m_i^2 \in F$, $m_i^3 = (m_i^3[1], m_i^3[2])$ where $\mathscr{T}_{-i}$-measurable $m_i^3[1] : T_{-i} \to \Delta(A)$ and $m_i^3[2] \in F$, $m_i^4 \in \Delta(A)$, $m_i^5 \in \{1, \ldots, n\}$, and $m_i^6 \in \mathbb{N} = \{1, 2, \ldots\}$. The only modification we need is to impose the measurability requirement over the message space. We set $\mathcal{T}_i = (T_i \times M_i, \mathscr{T}_i \times \mathscr{M}_i)$ where $\mathscr{M}_i = \mathscr{T}_i \times \mathscr{F} \times \prod_{f \in F} \mathscr{Y}_i[f] \times \mathscr{F} \times \mathscr{A} \times 2^N \times 2^{\mathbb{N}}$ is its associated sigma-algebra.

The rest of the proof is completed by appropriately adapting that of Theorem 4 to the current setup. ■

References

[1] Abreu, D. and H. Matsushima (1992), "Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information," Mimeo, Princeton University.

[2] Battigalli, P., and M. Siniscalchi (2003), "Rationalizable Bidding in First Price Auctions," *Games and Economic Behavior*, vol. 45, 38-72.

[3] Bergemann, D. and S. Morris (2008), "Interim Rationalizable Implementation," Mimeo.

[4] Bergemann, D. and S. Morris (2011), "Robust Implementation in General Mechanisms," *Games and Economic Behavior*, vol. 71, 261-281.

[5] Bergemann, D., S. Morris, and O. Tercieux (2011), "Rationalizable Implementation," *Journal of Economic Theory*, vol. 146, 1253-1274.

[6] Bernheim, D. (1984), "Rationalizable Strategic Behavior." *Econometrica* vol. 52, 1007-1028.

[7] Brandenburger, A. and E. Dekel (1987), "Rationalizability and Correlated Equilibria," *Econometrica* vol. 55, 1391-1402.

[8] Chen, Y-C., T. Kunimoto, and Y. Sun (2019), "Continuous Implementation with Small Transfers," *Working Paper*, Singapore Management University.

[9] Dasgupta, P., P. Hammond, and E. Maskin (1979), "Implementation of Social Choice Rules: Some General Results on Incentive Compatibility," *Review of Economic Studies* vol. 46, 195-216.

[10] D'Aspremont, C. and P.-A. Gerard-Varet (1979), "Incentives and Incomplete Information," *Journal of Public Economics* vol. 11, 25-45.

[11] Dekel, E., D. Fudenberg, and S. Morris (2007), "Interim Correlated Rationalizability," *Theoretical Economics*, vol. 2, 15-40.

[12] Green, J. R. and J.-J. Laffont (1979), *Incentives in Public Decision Making*, Amsterdam, North Holland.

[13] Harris, M. and R. Townsend (1981), "Resource Allocation with Asymmetric Information," *Econometrica* vol. 49, 33-64.

[14] Holmström, B. and R. B. Myerson (1983),, "Efficient and Durable Decision Rules with Incomplete Information," *Econometrica* vol. 51, 1799-1819.

[15] Jackson, M. (1991), "Bayesian Implementation," *Econometrica*, vol. 59, 461-477.

[16] Jehiel, P. and B. Moldovanu (2001), "Efficient Design with Interdependent Valuations," *Econometrica*, Vol. 69, 1237-1259

[17] Krishna, V. and M. Perry (2000), "Efficient Mechanism Design," *Working Paper*, Penn State University.

[18] Kunimoto, T. (2019), "Mixed Bayesian Implementation in General Environments," *Journal of Mathematical Economics*, vol. 82, 247-263.

[19] Kunimoto, T. and R. Serrano (2019), "Rationalizable Implementation of Correspondences," *Mathematics of Operations Research*, vol. 44, 1326-1344.

[20] Lipman, B. (1994), "A Note on the Implications of Common Knowledge of Rationality," *Games and Economic Behavior*, vol. 6, 114-129.

[21] Maskin, E. (1999), "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies*, vol. 66, 23-38.

[22] Myerson, R. B. (1979), "Incentive Compatibility and the Bargaining Problem," *Econometrica* vol. 47, 61-73.

[23] Myerson, R. B. (1981), "Optimal Auction Design," *Mathematics of Operations Research* vol. 6, 58-73.

[24] Oury, M. and O. Tercieux (2012), "Continuous Implementation," *Econometrica*, vol. 80, 1605-1637.

[25] Palfrey, T. and S. Srivastava (1989), "Implementation with Incomplete Information in Exchange Economies," *Econometrica*, vol. 57, 115-134.

[26] Pearce, D. (1984), "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* vol. 52, 1029-1050.

[27] Postlewaite, A. and D. Schmeidler (1986), "Implementation in Differential Information Economies," *Journal of Economic Theory* vol. 39, 14-33.

[28] Serrano, R. and R. Vohra (2001), "Some Limitations of Virtual Bayesian Implementation," *Econometrica*, vol. 69, 785-792.

[29] Wilson, R. (1978), "Information, Efficiency and the Core of an Economy," *Econometrica* vol. 46, 807-816.