Robust Implementation in Rationalizable Strategies in General Mechanisms*

Takashi Kunimoto† Rene Saran‡

May 18, 2021

Abstract

A social choice function (SCF) is robustly implementable in rationalizable strategies (RoRat-implementable) if every rationalizable strategy profile on every type space results in outcomes consistent with it. First, we establish the equivalence between RoRat-implementation and "weak rationalizable implementation", the latter being a "type-free" concept. Second, using the equivalence result, we identify weak robust monotonicity as a necessary and almost sufficient condition for RoRat-implementation. This exhibits a contrast with robust implementation in interim equilibria (RoEq-implementation), i.e., every equilibrium on every type space must achieve outcomes consistent with the SCF. Bergemann and Morris (2011) show that strict robust monotonicity is a necessary and almost sufficient condition for RoEq-implementation. We argue that strict robust monotonicity is strictly stronger than weak robust monotonicity, which further implies that, within general mechanisms, RoRat-implementation is more permissive than RoEq-implementation. The gap between RoRat-implementation and RoEq-implementation stems from the strictly stronger nonemptiness requirement inherent in the latter concept.

JEL: C72; D78; D80

Keywords: Ex post incentive compatibility, rationalizability, interim equilibrium, robust implementation, weak rationalizable implementation, weak robust monotonicity

*We are grateful to the editor, an advisory editor, and two anonymous referees for helpful comments that have significantly improved the paper. We thank Roberto Serrano for helpful comments.

†School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; tkunimoto@smu.edu.sg

‡Department of Economics, University of Cincinnati, 2906 Woodside Dr, Cincinnati, Ohio 45221, USA. *Email*: rene.saran@uc.edu; *Tel:* +1 513 556 1528

1 Introduction

We consider robust (full) implementation of a social choice function (SCF) in (interim correlated) rationalizable strategies (henceforth, RoRat-implementation). That is, we want the designer to construct a mechanism such that, *regardless* of the type space, *all* rationalizable outcomes are consistent with the SCF. We thus take a global approach to robustness, following the seminal work of Bergemann and Morris (henceforth, BM, 2005, 2009, 2011). However, we depart from BM as they assume interim equilibrium whereas we assume rationalizability as the solution concept. Specifically, we benchmark our work against BM (2011), which analyzes robust implementation of an SCF in interim equilibria (henceforth, RoEq-implementation).

Rationalizability characterizes outcomes that are consistent with common certainty of rationality (Dekel et al., 2007). On any given type space, rationalizability is a weaker solution concept than interim equilibrium. Interim equilibrium relies on the assumption of rational expectations, whereby all types of all players have correct beliefs about each other's strategies. Rationalizability, in contrast, is a set-valued concept that does not assume rational expectations, allowing different types of a player to hold distinct beliefs about other's strategies in order to rationalize their own behavior.

Fixing the solution concept (interim equilibrium or rationalizability), robust implementation imposes two requirements: (i) *Nonemptiness*: The solution concept must be nonempty on every type space and (ii) *Uniqueness*: Every outcome generated by the solution concept on every type space must be consistent with the SCF. As rationalizability is weaker than interim equilibrium on every type space, the nonemptiness requirement in RoRat-implementation is weaker than that in RoEq-implementation. At first glance, we might then think that RoRat-implementation imposes a stronger uniqueness requirement than RoEq-implementation. That turns out not to be the case. Indeed, the uniqueness requirement in RoRat-implementation is the same as that in RoEq-implementation. This is because every rationalizable action on a given type space can be obtained in interim equilibrium on another type space (this result is known in the literature on epistemic foundations; see, for instance, Remark 2 in Dekel et al., 2007). Thus, *a priori*, RoRat-implementation is weaker than RoEq-implementation. But is it strictly weaker? If yes, then what is the precise difference between these robust implementation notions? We answer these questions in this paper.

We first show that RoRat-implementation is equivalent to weak rationalizable implementation (henceforth, wRat-Implementation), which is a "type-free" implementation concept defined in the appendix of BM (2010). This result parallels the *almost* equivalence between RoEq-implementation and rationalizable implementation (henceforth, Rat-Implementation).

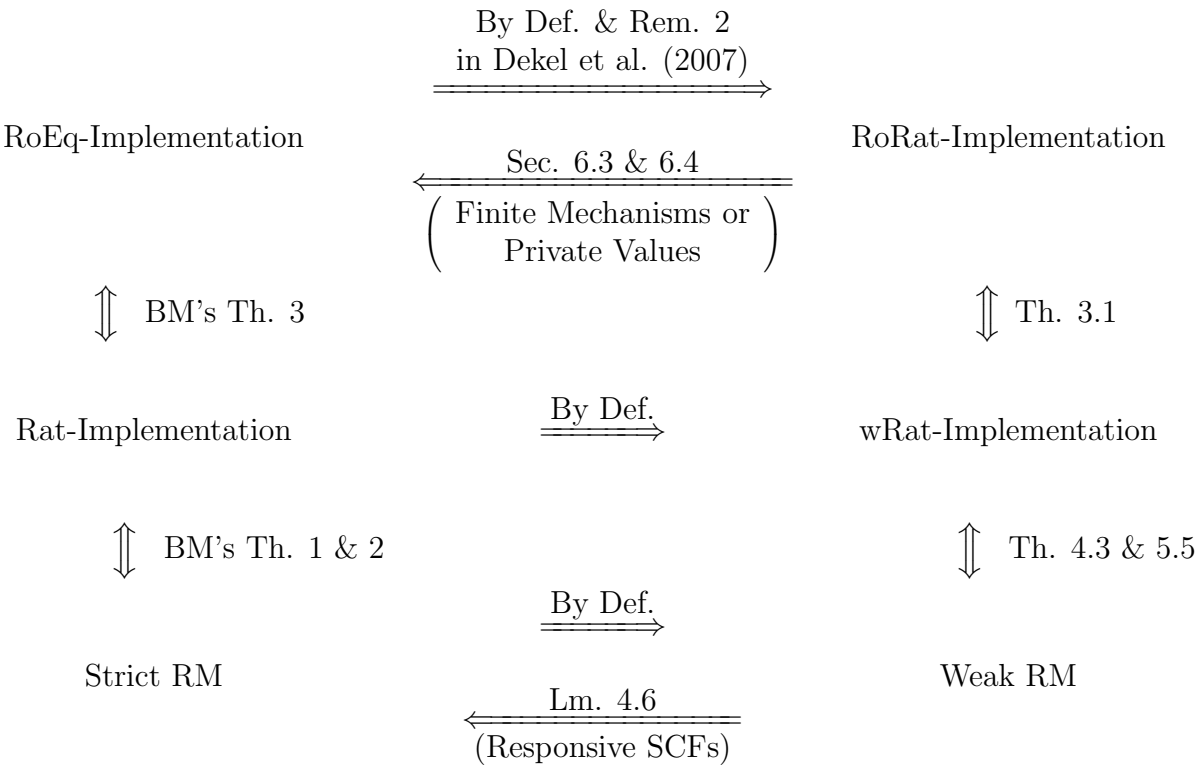

Figure 1: Summary of relationships between different implementation and monotonicity concepts. BM stands for Bergemann and Morris (2011); Rem. stands for Remark; Def. stands for Definition; Lm. stands for Lemma; Sec. stands for Section; and Th. stands for Theorem.

Rat-implementation is another "type-free" implementation concept – introduced in BM (2011) – that, as the name suggests, implies wRat-implementation. BM (2011) show that RoEq-implementation implies Rat-implementation, and the converse is true whenever the implementing mechanism has nonempty interim equilibria in all type spaces.

The equivalence between RoRat-implementation and wRat-implementation proves to be extremely useful on two counts:

(i) We use the equivalence to identify *weak robust monotonicity* (weak RM) as a necessary and almost sufficient condition for RoRat-implementing an SCF.

(ii) Following (i), we are able to provide a precise relationship between RoRat-implementation and RoEq-implementation. Thus, unlike many papers in the implementation literature that usually derive only necessary and sufficient conditions for different implementation notions, an important contribution of this paper is that we precisely establish the relationship between two different implementation notions, as described and summarized in Figure 1.

BM (2011) show that *strict* robust monotonicity (strict RM) is necessary and almost

sufficient for RoEq-implementation – as well as Rat-implementation. We show that strict RM implies weak RM and the converse is true for "responsive" SCFs but not more generally. In Example 6.1, we present an SCF that satisfies weak RM (and the other mild sufficient condition for RoRat-implementation) but not strict RM. Thus, there exist SCFs that are RoRat-implementable but not RoEq-implementable.

In light of the fact that the uniqueness requirements in the two implementation notions are the same, the gap between RoRat-implementation and RoEq-implementation is due to the strictly stronger nonemptiness requirement imposed in the latter notion. That is, it is possible that the SCF is RoRat-implementable but there exists a type space on which the implementing mechanism has no interim equilibria. The assumption of rational expectations plays a critical role in generating this possibility, as intuitively explained in Section 1.1. Although one might find it pathological that the induced game has no equilibria, the lack of equilibria in *some* type space does not preclude the existence of interim equilibria in other type spaces. Indeed, in Section 6.2 we show that the canonical mechanism that RoRat-implements the desired SCF has nonempty interim equilibria on type spaces that are typically found in the applied literature. Thus, pursuing RoRat-implementation seems to be a reasonable way of obtaining a more permissive result by relaxing the nonemptiness of equilibria in *all* type spaces.

If we restrict the designer to use finite mechanisms, then interim equilibria will exist on all type spaces. Hence, RoEq-implementation and RoRat-implementation are equivalent under this restriction. In other words, robustness consideration makes the difference between rationalizable strategies and equilibria moot within the class of finite mechanisms but not more generally. The designer can and must use countably infinite mechanisms if her aim is to RoRat-implement an SCF which cannot be RoEq-implemented. The implementation literature relies on countably infinite mechanisms to obtain tight necessary and sufficient conditions. In that spirit, we too construct a countably infinite mechanism to prove that weak RM is almost sufficient for RoRat-implementation. However, such constructions have been criticized for being impractical (see, for e.g., Jackson, 1992).

In the context of complete information environments, Bergemann et al. (2011) show that the necessary condition for implementation in rationalizable strategies is stronger than Maskin monotonicity, which is necessary and almost sufficient for Nash implementation (Maskin, 1999). In their Section 5, they also give an example of a Nash implementable SCF that is not implementable in rationalizable strategies. Recently, Xiong (2018) has provided a complete characterization of SCFs that are implementable in rationalizable strategies. The implementing mechanism in Xiong (2018) also Nash implements the SCF. Thus, in complete information environments, the designer can implement a strictly larger set of SCFs

in equilibrium than in rationalizable strategies.[1] In an interesting contrast, we show that, when it comes to robust implementation, the designer can RoRat-implement a strictly larger set of SCFs than those she can RoEq-implement.

1.1 Why RoRat-implementation might succeed where RoEq-implementation fails?

Consider an environment with three alternatives $\{a, b, c\}$. Any lottery ℓ on this set of alternatives can be represented by a tuple $(\ell[a], \ell[b])$ such that $\ell[a], \ell[b] \geq 0$ and $\ell[a] + \ell[b] \leq 1$. Graphically, this defines a triangle, as shown in Figure 2.

Suppose there is some player i with three payoff types, say, θ_i, θ_i', and θ_i''. Furthermore, we assume that the SCF f is "non-responsive" to θ_i' and θ_i'', i.e., for all types of the other players, the SCF prescribes the same outcome whether i's type is θ_i' or θ_i''. Then, in particular, $f(\theta_i', \theta_{-i}') = f(\theta_i'', \theta_{-i}')$ for some θ_{-i}', as shown in Figure 2. Also, suppose that the SCF f is "responsive" to θ_i and θ_i' such that $f(\theta_i, \theta_{-i}') \neq f(\theta_i', \theta_{-i}')$, as in Figure 2.

Throughout this discussion, we fix some type space $\mathcal{T}$ such that it includes four types of player i: $\hat{t}_i$, t_i, t_i', and t_i''. Type $\hat{t}_i$ has the payoff type θ_i and believes that the opponents' type profile is some $\hat{t}_{-i}$ with the corresponding payoff-type profile as some θ_{-i}. Types t_i, t_i', and t_i'' are such that their corresponding payoff types are θ_i, θ_i', and θ_i'', respectively, and all the three types believe that the opponents' type profile is some t_{-i}' with the corresponding payoff-type profile as some θ_{-i}'.

Suppose there were a mechanism that RoEq-implemented f. Let g be the outcome function of the mechanism, mapping message profiles to lotteries. Pick a (pure) interim equilibrium σ of the mechanism on the type space $\mathcal{T}$. Then, by definition of RoEq-implementation, $g(\sigma(\hat{t})) = f(\theta)$ whereas $g(\sigma(t')) = f(\theta') = f(\theta_i'', \theta_{-i}') = g(\sigma_i(t_i''), \sigma_{-i}(t_{-i}'))$. If instead of reporting their equilibrium messages $\sigma(\hat{t})$, the types $\hat{t}$ of the players were to jointly misreport their messages as $\sigma(t')$, then that will implement the outcome $f(\theta')$, which is not desired under the type profile $\hat{t}$. Thus, the mechanism must incentivize some player to blow the whistle when the players jointly misreport in this manner. We show that it is impossible to incentivize player i of type $\hat{t}_i$ to be the whistle blower. In order to incentivize $\hat{t}_i$, the mechanism must offer her a deviation that generates an outcome to the right of the indifference curve passing through $f(\theta')$ in the payoff state θ – this indifference curve is given by the dashed-line labelled as $u_i = u_i\big(f(\theta_i', \theta_{-i}'), (\theta_i, \theta_{-i})\big)$ in Figure 2. But the region to the right of the dashed-line includes the set of all lotteries that are either better than $f(\theta') = g(\sigma(t'))$ for type t_i' –

[1] This is true only for SCFs. For multi-valued social choice correspondences, implementation in rationalizable strategies is strictly weaker than Nash implementation, as shown in Kunimoto and Serrano (2019). Also see Jain (2021).

these are lotteries to the right of the indifference curve labelled as $u_i = u_i\big(f(\theta_i', \theta_{-i}'), (\theta_i', \theta_{-i}')\big)$ in Figure 2 – or better than $f(\theta_i'', \theta_{-i}') = g(\sigma_i(t_i''), \sigma_{-i}(t_{-i}'))$ for type t_i'' – these are lotteries to the right of the indifference curve labelled as $u_i = u_i\big(f(\theta_i'', \theta_{-i}'), (\theta_i'', \theta_{-i}')\big)$ in Figure 2. Thus, any deviation by $\hat{t}_i$ that undermines the joint misreport will also be improving for either type t_i' or t_i'' of player i when they believe that the opponents play $\sigma_{-i}(t_{-i}')$, contradicting the fact that σ is an interim equilibrium.

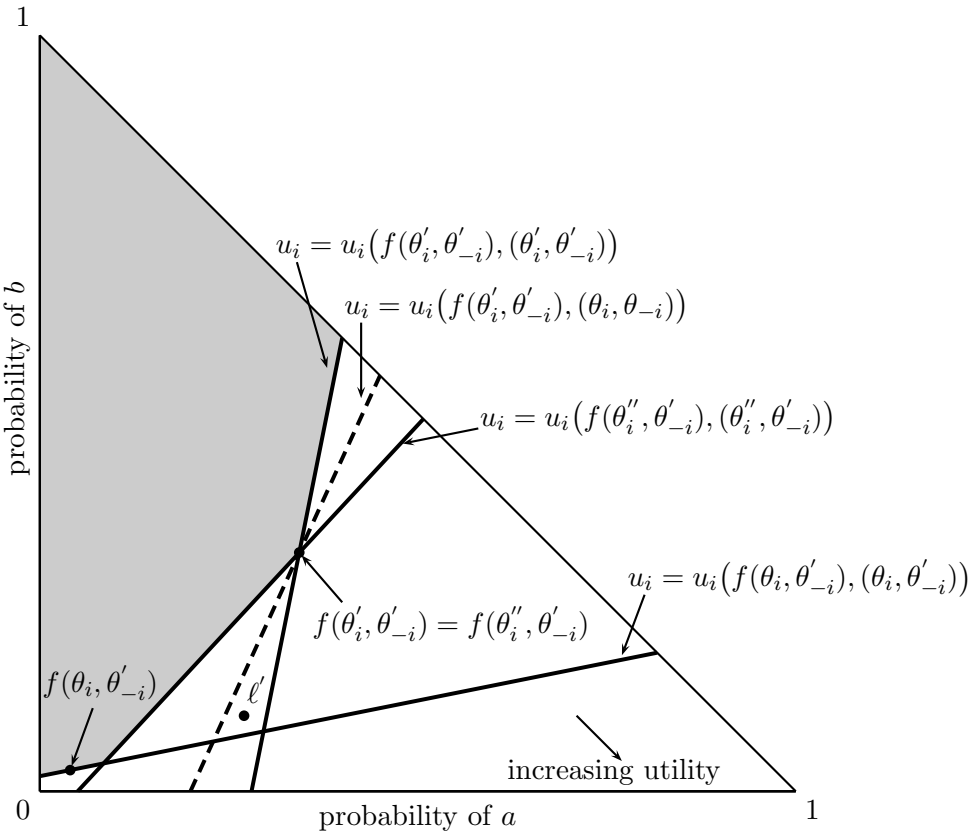

Figure 2: The case of non-responsive SCFs.

By playing their equilibrium messages against $\sigma_{-i}(t_{-i}')$, types t_i, t_i', and t_i'' obtain $f(\theta_i, \theta_{-i}')$, $f(\theta_i', \theta_{-i}')$, and $f(\theta_i'', \theta_{-i}')$, respectively. To sustain the equilibrium, any unilateral deviation against $\sigma_{-i}(t_{-i}')$ can only generate lotteries that lie in the *intersection* of the lower contour sets of these types at their respective equilibrium outcomes, which is given by the light-gray shaded region in Figure 2. But this region does not intersect the region of lotteries that incentivize $\hat{t}_i$ to blow the whistle on the joint misreport.

That unilateral deviations by a player from an interim equilibrium must generate outcomes that lie in the intersection of the lower contour sets of *all* types of that player at their respective equilibrium outcomes is due to rational expectations: In equilibrium, all types

of the player have the same belief about the strategies of her opponents. Rationalizability, in contrast, is a set-valued concept that does not assume rational expectations. Different types of a player can hold different beliefs about the strategies of the opponents in order to rationalize their own behavior. For instance, it is possible that three message profiles, say, m_{-i}, $m^{'}_{-i}$, and $m^{''}_{-i}$ are rationalizable at type profile $t^{'}_{-i}$, and type t_i has a rationalizable message m_i that is a best response to the belief that the opponents with type profile $t^{'}_{-i}$ play m_{-i}; type $t^{'}_i$ has a rationalizable message $m^{'}_i$ that is a best response to the belief that the opponents with type profile $t^{'}_{-i}$ play $m^{'}_{-i}$; and finally, type $t^{''}_i$ has a rationalizable message $m^{''}_i$ that is a best response to the belief that the opponents with type profile $t^{'}_{-i}$ play $m^{''}_{-i}$.

If the mechanism RoRat-implements the SCF f, then $g(m_i, m_{-i}) = f(\theta_i, \theta^{'}_{-i})$, $g(m^{'}_i, m^{'}_{-i}) = f(\theta^{'}_i, \theta^{'}_{-i})$, and $g(m^{''}_i, m^{''}_{-i}) = f(\theta^{''}_i, \theta^{'}_{-i})$. Now it becomes possible to undermine joint misreports of the kind discussed earlier without compromising the incentives of the types t_i, $t^{'}_i$, and $t^{''}_i$ of player i to play their respective rationalizable messages. To see this, suppose types $\hat{t}$ of the players were to jointly misreport their messages as $(m^{'}_i, m^{'}_{-i})$. Then, that will implement the outcome $f(\theta^{'})$, which is not desired under the type profile $\hat{t}$. The mechanism can incentivize type $\hat{t}_i$ to blow the whistle on this joint misreport by offering a deviation that generates the lottery $\ell^{'}$ when the opponents play $m^{'}_{-i}$. Since the lottery $\ell^{'}$ is worse than $f(\theta^{'})$ for type $t^{'}_i$, the message $m^{'}_i$ remains rationalizable for her against $m^{'}_{-i}$. At the same time, offering this deviation against $m^{'}_{-i}$ does not change the incentives of types t_i and $t^{''}_i$ to play their respective rationalizable messages, m_i and $m^{''}_i$. It thus follows that by relaxing rational expectations, RoRat-implementation might succeed where RoEq-implementation fails.

We should emphasize that the argument for the failure of RoEq-implementation made above hinges critically on the assumption that the SCF is non-responsive. The same argument cannot be extended to the case when the SCF is responsive to $\theta^{'}_i$ and $\theta^{''}_i$ such that $f(\theta^{'}_i, \theta^{'}_{-i}) \neq f(\theta^{''}_i, \theta^{'}_{-i})$. See Figure 3, which is the same as Figure 2 except for the change in the lottery $f(\theta^{''}_i, \theta^{'}_{-i})$. In that case, if there were a mechanism that RoEq-implemented the SCF f and σ were an equilibrium of the mechanism on the type space $\mathcal{T}$, then type $t^{''}_i$ must strictly prefer $f(\theta^{''}_i, \theta^{'}_{-i})$ to both $f(\theta_i, \theta^{'}_{-i})$ and $f(\theta^{'}_i, \theta^{'}_{-i})$. This is because type $t^{''}_i$ can generate the latter two outcomes by unilaterally deviating to either $\sigma_i(t_i)$ or $\sigma_i(t^{'}_i)$ when the opponents are playing $\sigma_{-i}(t^{'}_{-i})$. As a result, the intersection of the lower contour sets of the three types t_i, $t^{'}_i$, and $t^{''}_i$ at their respective equilibrium outcomes will overlap with the region of lotteries that incentivize $\hat{t}_i$ to blow the whistle on the joint misreport $\sigma(t^{'})$ by types $\hat{t}$ (this overlap is given by the dark-gray shaded region in Figure 3). More generally, while the necessary condition for RoEq-implementation can be strictly stronger than the one for RoRat-implementation when it comes to non-responsive SCFs (Example 6.1), the two necessary conditions coincide for responsive SCFs (Lemma 4.6).

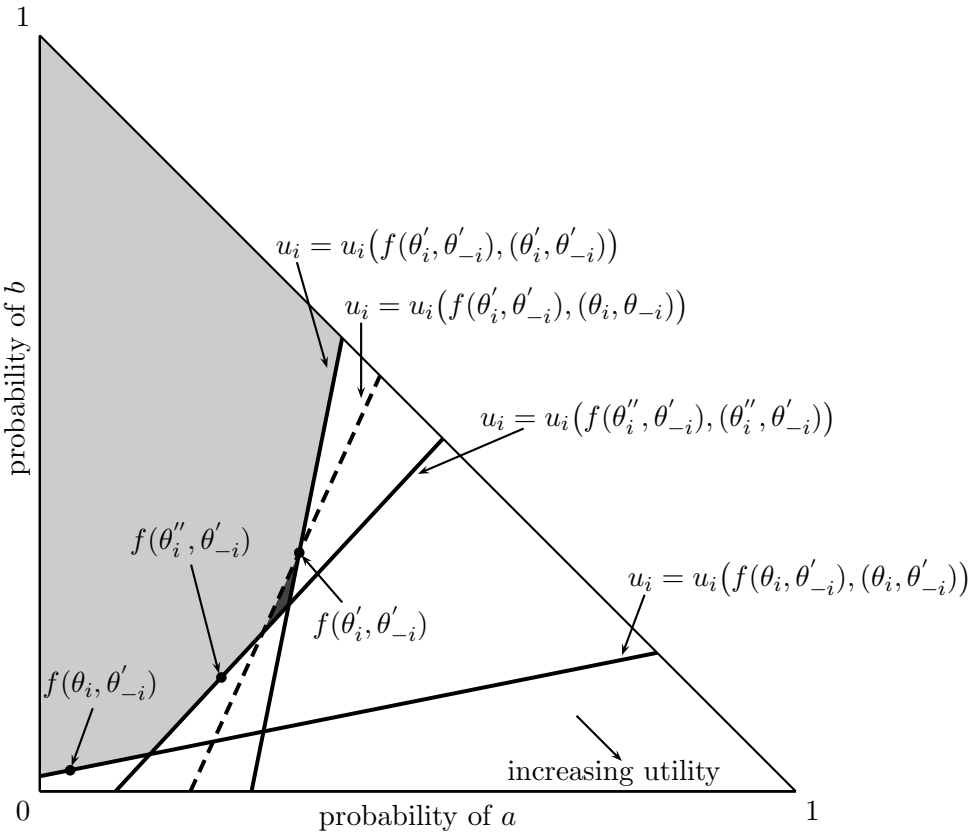

Figure 3: The case of responsive SCFs.

There are several economically relevant environments where the SCF is non-responsive. For example, in the context of voting, if there are two distinct payoff types of a player (viz., "extreme left" or "extreme right") such that the player is in the minority regardless of the payoff types of the opponents, then the Condorcet winner will not be responsive to those two payoff types of the player. As another example, suppose that the sum of the players' payoff types is either strictly greater or strictly less than a threshold in all payoff states and a public good is provided if and only if the sum of the players' payoff types is greater than the threshold. Then the decision to provide the public good will not be responsive to two sufficiently close payoff types of a player. As a final example, suppose the SCF is Rawlsian, i.e., it chooses the alternative that maximizes the utility of the worst-off individual in each payoff state. If a player has a payoff type such that she is never the worst-off individual regardless of the payoff types of the opponents, then the SCF will not be responsive to an even "higher" payoff type of that player (i.e., a payoff type that leads to a higher utility for each alternative). Indeed, even the utilitarian SCF that chooses the alternative that maximizes the sum of individuals' utilities can be non-responsive (see Example 6.1). Thus,

there is a rich set of economically relevant environments in which it might be feasible to RoRat-implement an SCF that is not RoEq-implementable.

The rest of the paper is organized as follows. We present the preliminary definitions in Section 2. In Section 3, we show that RoRat-implementation is equivalent to wRat-implementation. In Sections 4 and 5, respectively, we show that weak RM is necessary and almost sufficient for RoRat-implementation. We compare RoRat-implementation and RoEq-Implementation in Section 6 before concluding in Section 7. The Appendix contains the proofs omitted from the main body of the paper.

2 Preliminaries

There is a finite set of players $I = \{1, \ldots, n\}$. A player's *payoff type* is $\theta_i \in \Theta_i$, where we assume that Θ_i is finite. A *payoff state* is $\theta \in \Theta = \times_{i \in N} \Theta_i$. Denote $\Theta_{-i} \equiv \Theta_1 \times \cdots \times \Theta_{i-1} \times \Theta_{i+1} \times \cdots \times \Theta_n$.[2] There is a countable set of alternatives A with at least two elements. We let $\Delta(A)$ to be the set of lotteries over A.[3] We denote an arbitrary lottery by ℓ, and let a be the lottery that puts probability 1 on alternative a. For any lottery ℓ, let $\ell[a]$ be the probability assigned by ℓ to $a \in A$. Let $\mathbb{Z}$ be any countable set of indices. For any countable set of lotteries $\{\ell_z\}_{z \in \mathbb{Z}}$ and corresponding weights $\{\alpha_z\}_{z \in \mathbb{Z}}$ such that $\alpha_z \geq 0, \forall z$, and $\sum_{z \in \mathbb{Z}} \alpha_z = 1$, we let $\sum_{z \in \mathbb{Z}} \alpha_z \ell_z$ be the lottery that is obtained as a reduced form of the compound lottery in which for all $z \in \mathbb{Z}$, lottery ℓ_z is selected with probability α_z.

We endow A with the discrete topology. Thus, A is separable and completely metrizable by the discrete metric, and hence it is a Polish space. As a result, $\Delta(A)$ is also Polish under the weak* topology (Aliprantis and Border, 2006, Theorem 15.15). Therefore, $\Delta(A)$ contains a countable dense subset, which we denote by $\Delta^*(A)$.

Preferences of player i over the set of lotteries are represented by the von Neumann-Morgenstern expected utility function $u_i : \Delta(A) \times \Theta \to \Re$. Thus, for any payoff state θ and lottery ℓ, $u_i(\ell, \theta) = \sum_{a \in A} \ell[a] u_i(a, \theta)$. We assume that utilities are bounded to ensure that expected utility is well defined over the space of lotteries with countable support, i.e., for all $i \in I$ and $\theta \in \Theta$, there exists $\zeta > 0$ such that $|u_i(\ell, \theta)| \leq \zeta$ for all $\ell \in \Delta(A)$.[4]

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

[3] For any set X, we will use $\Delta(X)$ to denote the set of probability measures over X.

[4] See Blackwell and Girshick (1954) for an axiomatization of expected utility over all discrete probability measures on a set, which results in bounded utilities.

2.1 Type Space

A type space is a collection $\mathcal{T} = (T_i, \hat{\theta}_i, \hat{\pi}_i)_{i \in I}$ such that for each $i \in I$, T_i is countable, $\hat{\theta}_i : T_i \to \Theta_i$ and $\hat{\pi}_i : T_i \to \Delta(T_{-i})$. A player's *type* $t_i \in T_i$ defines her *payoff type* $\hat{\theta}_i(t_i) \in \Theta_i$ and her *belief type* $\hat{\pi}_i(t_i) \in \Delta(T_{-i})$. For any $t_{-i} \in T_{-i}$, we let $\hat{\pi}_i(t_i)[t_{-i}]$ denote the probability that player i of type t_i assigns to other players having types t_{-i}. We assume that $\hat{\theta}_i : T_i \to \Theta_i$ is surjective for all $i \in I$, i.e., no payoff type is redundant.

For each $i \in I$, let $Z_i^1 = \Delta(\Theta_{-i})$ be the set of all possible beliefs that player i can have about the payoff types of the other agents.

Given the type space $\mathcal{T}$, for each player i and type $t_i \in T_i$, we let $z_i^1(t_i) \in Z_i^1$ be the *first-order belief* of t_i, i.e., $z_i^1(t_i)[\theta_{-i}] = \sum_{t_{-i} \in T_{-i} : \hat{\theta}_{-i}(t_{-i}) = \theta_{-i}} \hat{\pi}_i(t_i)[t_{-i}]$ for all $\theta_{-i} \in \Theta_{-i}$.

2.2 Social Choice Function and Mechanism

The planner's objective is specified by a *social choice function (henceforth, SCF)* as a function $f : \Theta \to \Delta(A)$.

We say that the SCF f is *responsive to* θ_i *and* θ_i', denoted by $\theta_i' \not\sim_i^f \theta_i$, if $f(\theta_i, \theta_{-i}) \neq f(\theta_i', \theta_{-i})$ for some $\theta_{-i} \in \Theta_{-i}$. Otherwise, f is *non-responsive to* θ_i *and* θ_i', denoted by $\theta_i' \sim_i^f \theta_i$.

The SCF f is *responsive* if for all $i \in I$ and $\theta_i, \theta_i' \in \Theta_i$: $\theta_i \neq \theta_i' \Rightarrow \theta_i \not\sim_i^f \theta_i'$. Otherwise, f is *non-responsive*.

A *mechanism* $\Gamma = ((M_i)_{i \in I}, g)$, where M_i is a countable nonempty set of messages for player i, $M = \times_{i \in I} M_i$, and $g : M \to \Delta(A)$ is the outcome function. The mechanism $\Gamma = ((M_i)_{i \in I}, g)$ is *finite* if M_i is finite for all $i \in I$.

2.3 Rationalizable Strategies

Fix a type space $\mathcal{T}$ and mechanism $\Gamma = ((M_i)_{i \in I}, g)$. A *message correspondence profile* $S = (S_1, \ldots, S_n)$, where each $S_i : T_i \to 2^{M_i}$.

Let $\mathbb{S}$ be the collection of all such message correspondence profiles. The collection $\mathbb{S}$ is a complete lattice with the natural ordering of set inclusion: $S \leq S'$ if $S_i(t_i) \subseteq S_i'(t_i)$ for all $i \in I$ 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 $i \in I$ 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 $i \in I$ and $t_i \in T_i$.

We define the *best response operator* $b : \mathbb{S} \to \mathbb{S}$ as follows:

$$b_i(S)[t_i] \equiv \left\{ m_i \in M_i : \begin{array}{ll} & \exists \lambda_i \in \Delta(T_{-i} \times M_{-i}) \text{ such that} \\ \text{(i)} & m_i \in \arg\max_{m_i' \in M_i} \sum_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}), \hat{\theta}(t_i, t_{-i})) \\ \text{(ii)} & \text{marg}_{T_{-i}} \lambda_i = \hat{\pi}_i(t_i) \\ \text{(iii)} & \lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}(t_{-i}) \end{array} \right\},$$

where $S_{-i}(t_{-i}) = \times_{j \neq i} S_j(t_j)$ for each $t_{-i} \in T_{-i}$.

Observe that b is increasing by definition: i.e., $S \leq S' \Rightarrow b(S) \leq b(S')$. Since b is increasing and $\mathbb{S}$ is a complete lattice, by Tarski's fixed point theorem, there is a largest fixed point of b, which we label B^∞. Thus, (i) $b(B^\infty) = B^\infty$ and (ii) $b(S) \geq S \Rightarrow S \leq B^\infty$.

B^∞ is the (interim correlated) *rationalizable* message correspondence profile (Dekel et al., 2007). For each type of each player, it characterizes the messages that are consistent with common certainty of rationality.

2.4 $\mathcal{S}^\infty$ Correspondence

As we will see, insisting on implementation that is robust to the underlying type space will force the solution concept to depend only on the payoff types of the individuals. Hence, we need to define strategies that are "rationalizable" for payoff types.

Fix a mechanism $\Gamma = ((M_i)_{i \in I}, g)$. A *message correspondence profile with payoff-type domain* $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_n)$, where each $\mathcal{S}_i : \Theta_i \to 2^{M_i}$.

Let $\mathbb{S}^\Theta$ be the collection of such message correspondence profiles with payoff-type domain. The collection $\mathbb{S}^\Theta$ is a complete lattice with the natural ordering of set inclusion: $\mathcal{S} \leq \mathcal{S}'$ if $\mathcal{S}_i(\theta_i) \subseteq \mathcal{S}_i'(\theta_i)$ for all $i \in I$ and $\theta_i \in \Theta_i$. The largest element is $\bar{\mathcal{S}} = (\bar{\mathcal{S}}_1, \ldots, \bar{\mathcal{S}}_n)$, where $\bar{\mathcal{S}}_i(\theta_i) = M_i$ for each $i \in I$ and $\theta_i \in \Theta_i$. The smallest element is $\underline{\mathcal{S}} = (\underline{\mathcal{S}}_1, \ldots, \underline{\mathcal{S}}_n)$, where $\underline{\mathcal{S}}_i(\theta_i) = \emptyset$ for each $i \in I$ and $\theta_i \in \Theta_i$.

We define the *best response operator for payoff types* $b^\Theta : \mathbb{S}^\Theta \to \mathbb{S}^\Theta$ as follows:

$$b_i^\Theta(\mathcal{S})[\theta_i] \equiv \left\{ m_i \in M_i : \begin{array}{ll} & \exists \psi_i \in (\Theta_{-i} \times M_{-i}) \text{ such that} \\ \text{(i)} & m_i \in \arg\max_{m_i'} \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i(g(m_i', m_{-i}), (\theta_i, \theta_{-i})) \\ \text{(ii)} & \psi_i(\theta_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in \mathcal{S}_{-i}(\theta_{-i}) \end{array} \right\},$$

where $\mathcal{S}_{-i}(\theta_{-i}) = \times_{j \neq i} \mathcal{S}_j(\theta_j)$ for each $\theta_{-i} \in \Theta_{-i}$.

As the operator b^Θ is increasing and $\mathbb{S}^\Theta$ is a complete lattice, by Tarski's fixed point theorem, there is a largest fixed point of b^Θ, which we denote by $\mathcal{S}^\infty$. Thus, (i) $b^\Theta(\mathcal{S}^\infty) = \mathcal{S}^\infty$

and (ii) $b^{\Theta}(\mathcal{S}) \geq \mathcal{S} \Rightarrow \mathcal{S} \leq \mathcal{S}^{\infty}$.

2.5 Notions of Implementation

In this section, we present four notions of implementation. The first one is robust implementation in rationalizable strategies (RoRat-implementation), which is the focus of this paper. BM (2011) define the next two: robust implementation in interim equilibria (RoEq-implementation) and rationalizable implementation (Rat-implementation). Finally, BM (2010) define weak rationalizable implementation (wRat-implementation).

2.5.1 RoRat-Implementation

To define RoRat-implementation, we start by defining what we mean by implementation in rationalizable strategies on a specific type space.

Definition 2.1. A mechanism $\Gamma = ((M_i)_{i\in I}, g)$ *implements the SCF f in rationalizable strategies on the type space $\mathcal{T}$* if, for all $t \in T$, we have

$$\text{(nonemptiness) } B^{\infty}(t) \neq \emptyset \quad \text{and} \quad \text{(uniqueness) } g(m) = f(\hat{\theta}(t)), \forall m \in B^{\infty}(t).$$

We now define RoRat-implementation as implementation in rationalizable strategies over "all type spaces".

Definition 2.2. A mechanism Γ *robustly implements the SCF f in rationalizable strategies (or, RoRat-implements the SCF f)* if, for all type spaces $\mathcal{T}$, the mechanism implements f in rationalizable strategies on $\mathcal{T}$. The SCF f is *robustly implementable in rationalizable strategies (or, RoRat-implementable)* if there exists a mechanism that RoRat-implements f.

2.5.2 RoEq-Implementation

To define RoEq-implementation, consider a type space $\mathcal{T}$ and a mechanism $\Gamma = ((M_i)_{i\in I}, g)$. The resulting incomplete information game is denoted by $(\mathcal{T}, \Gamma)$. A *strategy for individual i* in this game is a mapping $\sigma_i : T_i \to \Delta(M_i)$. A strategy profile $\sigma = (\sigma_1, \ldots, \sigma_n)$ is an *interim equilibrium* of the game $(\mathcal{T}, \Gamma)$ if, for all $i \in I$, $t_i \in T_i$, and $m_i \in M_i$ with $\sigma_i(t_i)[m_i] > 0$, we have

$$\begin{aligned}
&\sum_{t_{-i}\in T_{-i}} \hat{\pi}_i(t_i)[t_{-i}] \sum_{m_{-i}\in M_{-i}} \sigma_{-i}(t_{-i})[m_{-i}] u_i\big(g(m_i, m_{-i}), \hat{\theta}(t_i, t_{-i})\big) \\
&\geq \sum_{t_{-i}\in T_{-i}} \hat{\pi}_i(t_i)[t_{-i}] \sum_{m_{-i}\in M_{-i}} \sigma_{-i}(t_{-i})[m_{-i}] u_i\big(g(m_i', m_{-i}), \hat{\theta}(t_i, t_{-i})\big), \forall m_i' \in M_i.
\end{aligned}$$

We then have the following notion of interim implementation:

Definition 2.3. A mechanism $\Gamma = ((M_i)_{i\in I}, g)$ *interim implements the SCF f on the type space $\mathcal{T}$* if (nonemptiness) the game $(\mathcal{T}, \Gamma)$ has an interim equilibrium and (uniqueness) for every interim equilibrium σ of the game $(\mathcal{T}, \Gamma)$, if $\sigma(t)[m] > 0$, then $g(m) = f(\hat{\theta}(t))$.

RoEq-implementation is defined as interim implementation over "all type spaces".

Definition 2.4. A mechanism Γ *robustly implements the SCF f in interim equilibria (or RoEq-implements the SCF f)* if, for all type spaces $\mathcal{T}$, the mechanism Γ interim implements f on $\mathcal{T}$. The SCF f is *robustly implementable in interim equilibria (or RoEq-implementable)* if there exists a mechanism that RoEq-implements f.

2.5.3 Rat-Implementation

BM (2011) define Rat-implementation as a "type-free" implementation concept by imposing the uniqueness and nonemptiness requirements directly on the $\mathcal{S}^\infty$ correspondence.

Definition 2.5. A mechanism $\Gamma = ((M_i)_{i\in I}, g)$ *rationalizably implements (or Rat-implements) the SCF f* if

1. (uniqueness) $m \in \mathcal{S}^\infty(\theta) \Rightarrow g(m) = f(\theta)$; and
2. (nonemptiness) For each $i \in I$ and $z_i^1 \in Z_i^1$, there exists a belief $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$ such that:
   (a) $\arg\max_{m_i' \in M_i} \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(m_i', m_{-i}), (\theta_i, \theta_{-i})\big) \neq \emptyset$ for all $\theta_i \in \Theta_i$.
   (b) $\psi_i(\theta_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$.
   (c) $\text{marg}_{\Theta_{-i}} \psi_i = z_i^1$.

The SCF f is *Rat-implementable* if there exists a mechanism that Rat-implements f.

Part (1) of the definition, i.e., the uniqueness requirement, states that every message profile in S^∞ must lead to a socially desirable outcome. Part (2), i.e., the nonemptiness requirement, imposes a strong existence condition: For every belief that agent i may have over the payoff types of the other agents, there must exist a belief over their messages in S_{-i}^∞ such that agent i has a best response *whatever* be his payoff type.

BM (2011, Theorem 3) prove that if a mechanism RoEq-implements an SCF, then the same mechanism also Rat-implements the SCF. The converse is true whenever the mechanism that Rat-implements the SCF has nonempty interim equilibria in all type spaces. That will

be the case if the message correspondence $\mathcal{S}^\infty$ satisfies the *ex post best response property*. The property, as defined in BM (2011), requires that for all $i \in I$ and $\theta_i \in \Theta_i$, there exist a single message $m_i^* \in \mathcal{S}_i^\infty(\theta_i)$ such that

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

for all $\theta_{-i} \in \Theta_{-i}$ and $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$. With the ex post best response property, the message profile m^* is an ex-post equilibrium, which in turn guarantees the nonemptiness of interim equilibria in all type spaces.

2.5.4 wRat-Implementation

BM (2010) define weak rationalizable implementation (wRat-implementation) by weakening the nonemptiness requirement while maintaining the uniqueness requirement in Rat-implementation. Specifically, they weaken the nonemptiness requirement (Part (2) in Definition 2.5) by allowing the belief ψ_i to depend on the payoff type of individual i.

Definition 2.6. A mechanism $\Gamma = ((M_i)_{i\in I}, g)$ *weakly rationalizably implements (or wRat-implements) the SCF* f if

1. (uniqueness) $m \in \mathcal{S}^\infty(\theta) \Rightarrow g(m) = f(\theta)$; and
2. (nonemptiness) For each $i \in I$, $\theta_i \in \Theta_i$ and $z_i^1 \in Z_i^1$, there exists a belief $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$ such that:
   (a) $\arg\max_{m_i' \in M_i} \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(m_i', m_{-i}), (\theta_i, \theta_{-i})\big) \neq \emptyset$.
   (b) $\psi_i(\theta_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$.
   (c) $\text{marg}_{\Theta_{-i}} \psi_i = z_i^1$.

The SCF f is *wRat-implementable* if there exists a mechanism that wRat-implements f.

3 Equivalence between RoRat-Implementation and wRat-Implementation

We now establish that RoRat-implementation is equivalent to wRat-implementation because the former imposes the same conditions on $\mathcal{S}^\infty$ as the latter.

Theorem 3.1. *The SCF f is RoRat-implementable by the mechanism Γ if and only if f is wRat-implementable by the same mechanism Γ.*

Proof. We first prove the following lemma:

Lemma 3.2. *Consider any mechanism* Γ. *The message profile* $m \in \mathcal{S}^\infty(\theta)$ *if and only if there exists a type space* $\mathcal{T}$ *such that* $m \in \bigcup_{t \in T:\hat{\theta}(t)=\theta} B^\infty(t)$.

Proof. ($\Rightarrow$) BM (2011, Proposition 1) show that if $m \in \mathcal{S}^\infty(\theta)$, then there exist a type space $\mathcal{T}$, a pure-strategy interim equilibrium σ, and a type profile t such that $\sigma(t) = m$ and $\hat{\theta}(t) = \theta$. Therefore, $m \in B^\infty(t)$.

($\Leftarrow$) Consider any type space $\mathcal{T}$. Define the message correspondence profile with payoff-type domain $\hat{\mathcal{S}} = (\hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_n)$ such that for all $i \in I$,

$$\hat{\mathcal{S}}_i(\theta_i^{'}) = \bigcup_{t_i \in T_i:\hat{\theta}_i(t_i)=\theta_i^{'}} B_i^\infty(t_i), \forall \theta_i^{'} \in \Theta_i.$$

If $m_i^{'} \in \hat{\mathcal{S}}_i(\theta_i^{'})$, then there exists $t_i^{'} \in T_i$ such that $\hat{\theta}_i(t_i^{'}) = \theta_i^{'}$ and $m_i^{'} \in B_i^\infty(t_i^{'})$. Thus, there exists a belief $\lambda_i \in \Delta(T_{-i} \times M_{-i})$ such that

$$m_i^{'} \in \arg\max_{m_i^{''} \in M_i} \sum_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i(g(m_i^{''}, m_{-i}), \hat{\theta}(t_i^{'}, t_{-i})),$$

$\text{marg}_{T_{-i}} \lambda_i = \hat{\pi}_i(t_i^{'})$ and $\lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in B_{-i}^\infty(t_{-i})$.

Define $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$ as follows:

$$\psi_i(\theta_{-i}, m_{-i}) = \sum_{t_{-i} \in T_{-i}:\hat{\theta}_{-i}(t_{-i})=\theta_{-i}} \lambda_i(t_{-i}, m_{-i}), \forall \theta_{-i}, m_{-i}.$$

Then $\psi_i(\theta_{-i}, m_{-i}) > 0$ implies that $m_{-i} \in \bigcup_{t_{-i} \in T_{-i}:\hat{\theta}_{-i}(t_{-i})=\theta_{-i}} B_{-i}^\infty(t_{-i}) = \hat{\mathcal{S}}_{-i}(\theta_{-i})$. Moreover, by construction,

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

Thus, $m_i^{'} \in b_i^\Theta(\hat{\mathcal{S}})[\theta_i^{'}]$. Hence, $b^\Theta(\hat{\mathcal{S}}) \geq \hat{\mathcal{S}}$. Therefore, $\hat{\mathcal{S}} \leq \mathcal{S}^\infty$.

Now suppose there exist $m \in M$ and $\theta \in \Theta$ such that $m \in \bigcup_{t \in T:\hat{\theta}(t)=\theta} B^\infty(t)$. Then $m \in \hat{\mathcal{S}}(\theta)$, and hence $m \in \mathcal{S}^\infty(\theta)$. This completes the proof of the lemma. □

We prove the necessity part of Theorem 3.1 first.

Suppose the SCF f is RoRat-implementable by the mechanism Γ. Then the following is true for all type spaces $\mathcal{T}$: For all $t \in T$, we have

$$B^\infty(t) \neq \emptyset \quad \text{and} \quad g(m) = f(\hat{\theta}(t)), \forall m \in B^\infty(t).$$

Pick any $\theta \in \Theta$. If $m \in \mathcal{S}^\infty(\theta)$, then it follows from Lemma 3.2 that there exists a type space $\mathcal{T}$ such that $m \in \bigcup_{t \in T': \hat{\theta}(t) = \theta} B^\infty(t)$. Hence, $g(m) = f(\theta)$.

Next, pick any i, θ_i and z_i^1. For each $j \neq i$, pick any $z_j^1 \in Z_j^1$. Define the type space $\mathcal{T}$ such that (i) $T_j = \{t_j^{\tilde{\theta}_j} : \tilde{\theta}_j \in \Theta_j\}$ for all $j \in I$, and (ii) $\hat{\theta}_j(t_j^{\tilde{\theta}_j}) = \tilde{\theta}_j$ and $\hat{\pi}_j(t_j^{\tilde{\theta}_j})[t_{-j}^{\tilde{\theta}_{-j}}] = z_j^1(\tilde{\theta}_{-j})$ for all $t_{-j}^{\tilde{\theta}_{-j}} \in T_{-j}$ and $t_j^{\tilde{\theta}_j} \in T_j$.

By our hypothesis of RoRat-implementation, $B_i^\infty(t_i^{\theta_i}) \neq \emptyset$. Therefore, there exists $\lambda_i \in \Delta(T_{-i} \times M_{-i})$ such that

1. $\arg\max_{m_i'} \sum_{t_{-i}^{\theta_{-i}}, m_{-i}} \lambda_i(t_{-i}^{\theta_{-i}}, m_{-i}) u_i\big(g(m_i', m_{-i}), \hat{\theta}(t_i^{\theta_i}, t_{-i}^{\theta_{-i}})\big) \neq \emptyset$.
2. $\text{marg}_{T_{-i}} \lambda_i = \hat{\pi}_i(t_i^{\theta_i})$
3. $\lambda_i(t_{-i}^{\theta_{-i}}, m_{-i}) > 0 \Rightarrow m_{-i} \in B_{-i}^\infty(t_{-i}^{\theta_{-i}})$.

Define $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$ as follows: for any $\theta_{-i} \in \Theta_{-i}$ and $m_{-i} \in M_{-i}$,

$$\psi_i(\theta_{-i}, m_{-i}) = \lambda_i(t_{-i}^{\theta_{-i}}, m_{-i}).$$

Then $\psi_i(\theta_{-i}, m_{-i}) > 0$ implies that $m_{-i} \in B_{-i}^\infty(t_{-i}^{\theta_{-i}})$. It follows from Lemma 3.2 that $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$. Lastly, by construction, $\text{marg}_{\Theta_{-i}} \psi_i = z_i^1$ and

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

We prove the sufficiency part of Theorem 3.1 next.

Suppose that the SCF f is wRat-implementable by the mechanism Γ. Consider any type space $\mathcal{T}$. If $m \in B^\infty(t)$, then it follows from Lemma 3.2 that $m \in \mathcal{S}^\infty(\hat{\theta}(t))$. Hence, $g(m) = f(\hat{\theta}(t))$.

We now show that $B^\infty(t) \neq \emptyset$ for all $t \in T$. Define the message correspondence profile $\hat{S} = (\hat{S}_1, \ldots, \hat{S}_n)$ such that, for all $i \in I$ and $t_i \in T_i$,

$$\hat{S}_i(t_i) = \mathcal{S}_i^\infty(\hat{\theta}_i(t_i)).$$

Pick any type $t_i \in T_i$. By our hypothesis of wRat-implementability, there exists a belief $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$ such that

(a) $\arg \max_{m_i' \in M_i} \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(m_i', m_{-i}), (\hat{\theta}_i(t_i), \theta_{-i})\big) \neq \emptyset$.

(b) $\psi_i(\theta_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$.

(c) $\text{marg}_{\Theta_{-i}}\psi_i = z_i^1(t_i)$.

By the definition of $\mathcal{S}_i^\infty(\hat{\theta}_i(t_i))$, we have

$$\emptyset \neq \arg\max_{m_i' \in M_i} \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(m_i', m_{-i}), (\hat{\theta}_i(t_i), \theta_{-i})\big) \subseteq \mathcal{S}_i^\infty(\hat{\theta}_i(t_i)).$$

Since $\hat{S}_i(t_i) = \mathcal{S}_i^\infty(\hat{\theta}_i(t_i))$, we also have $\hat{S}_i(t_i) \neq \emptyset$.

We now show that $\hat{S}_i(t_i) \leq b_i(\hat{S})[t_i]$. Consider any message $\tilde{m}_i \in \hat{S}_i(t_i)$. By our hypothesis of wRat-implementability, we have that for any $\theta \in \Theta$, $m' \in \mathcal{S}^\infty(\theta) \Rightarrow g(m') = f(\theta)$. Since $\tilde{m}_i \in \mathcal{S}_i^\infty(\hat{\theta}_i(t_i))$ and $\psi_i(\theta_{-i}, m_{-i}) > 0$ implies $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$, by wRat-implementability, we have

$$\sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(\tilde{m}_i, m_{-i}), (\hat{\theta}_i(t_i), \theta_{-i})\big) = \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(f(\hat{\theta}_i(t_i), \theta_{-i}), (\hat{\theta}_i(t_i), \theta_{-i})\big).$$

Thus, either every message in $\hat{S}_i(t_i)$ is a best response to ψ_i or none of the messages in $\hat{S}_i(t_i)$ is a best response to ψ_i. But, as already argued,

$$\hat{S}_i(t_i) = \mathcal{S}_i^\infty(\hat{\theta}_i(t_i)) \supseteq \arg\max_{m_i' \in M_i} \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(m_i', m_{-i}), (\hat{\theta}_i(t_i), \theta_{-i})\big) \neq \emptyset.$$

Thus, every message in $\hat{S}_i(t_i)$ is a best response to ψ_i.

Now pick any $m_i \in \hat{S}_i(t_i)$. As argued above,

$$m_i \in \arg\max_{m_i' \in M_i} \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(m_i', m_{-i}), (\hat{\theta}_i(t_i), \theta_{-i})\big).$$

Define the belief $\lambda_i \in \Delta(T_{-i} \times M_{-i})$ such that for all $(t_{-i}, m_{-i}) \in T_{-i} \times M_{-i}$,

$$\lambda_i(t_{-i}, m_{-i}) = \begin{cases} \hat{\pi}_i(t_i)[t_{-i}] \left(\frac{\psi_i(\hat{\theta}_{-i}(t_{-i}), m_{-i})}{z_i^1(t_i)[\hat{\theta}_{-i}(t_{-i})]}\right), & \text{if } \hat{\pi}_i(t_i)[t_{-i}] > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Since $\sum_{m_{-i}} \psi_i(\hat{\theta}_{-i}(t_{-i}), m_{-i}) = z_i^1(t_i)[\hat{\theta}_{-i}(t_{-i})]$, we have $\text{marg}_{T_{-i}}\lambda_i = \hat{\pi}_i(t_i)$. Moreover,

$$\lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow \psi_i(\hat{\theta}_{-i}(t_{-i}), m_{-i}) > 0 \Rightarrow m_{-i} \in \mathcal{S}_{-i}^\infty(\hat{\theta}_{-i}(t_{-i})) = \hat{S}_{-i}(t_{-i}).$$

Finally, for all $m_i' \in M_i$,

$$\sum_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i\big(g(m_i', m_{-i}), \hat{\theta}(t_i, t_{-i})\big)$$

$$= \sum_{\theta_{-i},m_{-i}} \left(\sum_{t_{-i}\in T_{-i}:\hat{\theta}_{-i}(t_{-i})=\theta_{-i}} \hat{\pi}_i(t_i)[t_{-i}] \frac{\psi_i(\theta_{-i},m_{-i})}{z_i^1(t_i)(\theta_{-i})} u_i\big(g(m_i^{'},m_{-i}),(\hat{\theta}_i(t_i),\theta_{-i})\big) \right)$$
$$= \sum_{\theta_{-i},m_{-i}} \psi_i(\theta_{-i},m_{-i}) u_i\big(g(m_i^{'},m_{-i}),(\hat{\theta}_i(t_i),\theta_{-i})\big),$$

where the last equality follows because $\sum_{t_{-i}\in T_{-i}:\hat{\theta}_{-i}(t_{-i})=\theta_{-i}} \hat{\pi}_i(t_i)[t_{-i}] = z_i^1(t_i)(\theta_{-i})$. Hence, we must have

$$m_i \in \arg\max_{m_i^{'}\in M_i} \sum_{t_{-i},m_{-i}} \lambda_i(t_{-i},m_{-i}) u_i\big(g(m_i^{'},m_{-i}),\hat{\theta}(t_i,t_{-i})\big).$$

We thus conclude that $m_i \in b_i(\hat{S})[t_i]$.

As $b(\hat{S}) \geq \hat{S}$, we have $\hat{S} \leq B^{\infty}$. Pick any $t \in T$. Then $B^{\infty}(t) \neq \emptyset$ because, as already shown, $\hat{S}(t) \neq \emptyset$. This completes the proof of the theorem. □

4 Necessary Condition

We now apply the equivalence result presented in the previous section to the establish the necessary condition for RoRat-implementation.

A *deception* is a profile of correspondences $\beta = (\beta_1, \ldots, \beta_n)$ such that $\beta_i : \Theta_i \to 2^{\Theta_i} \setminus \emptyset$ and $\theta_i \in \beta_i(\theta_i)$ for all $\theta_i \in \Theta_i$ and $i \in I$. A deception β is *unacceptable* if there exist $\theta \in \Theta$ and $\theta^{'} \in \beta(\theta)$ for which $f(\theta) \neq f(\theta^{'})$; otherwise, β is *acceptable*.

For each $i \in I$ and $\theta_i \in \Theta_i$, define

$$Y_i[\theta_i] \equiv \left\{ y : \Theta_{-i} \to \Delta(A) : \begin{array}{ll} \forall \theta_{-i} \in \Theta_{-i}, & \\ \text{either} & y(\theta_{-i}) = f(\theta_i,\theta_{-i}) \\ \text{or} & u_i\big(f(\theta_i,\theta_{-i}),(\theta_i,\theta_{-i})\big) > u_i\big(y(\theta_{-i}),(\theta_i,\theta_{-i})\big) \end{array} \right\}.$$

Thus, $Y_i[\theta_i]$ is the collection of all mappings $y : \Theta_{-i} \to \Delta(A)$ such that for every $\theta_{-i} \in \Theta_{-i}$, the lottery $y(\theta_{-i})$ is either equal to $f(\theta_i,\theta_{-i})$ or strictly worse than $f(\theta_i,\theta_{-i})$ for individual i in state (θ_i,θ_{-i}).

Definition 4.1. We say that an unacceptable deception β is *weakly refutable* if there exist $i \in I$, $\theta_i \in \Theta_i$, and $\theta_i^{'} \in \beta_i(\theta_i)$ satisfying $\theta_i^{'} \not\sim_i^f \theta_i$ such that for all $\tilde{\theta}_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ satisfying $\psi_i(\theta_{-i},\theta_{-i}^{'}) > 0 \Rightarrow \theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$, there exists $y \in Y_i[\tilde{\theta}_i]$ such that

$$\sum_{\theta_{-i},\theta_{-i}^{'}} \psi_i(\theta_{-i},\theta_{-i}^{'}) u_i\big(y(\theta_{-i}^{'}),(\theta_i,\theta_{-i})\big) > \sum_{\theta_{-i},\theta_{-i}^{'}} \psi_i(\theta_{-i},\theta_{-i}^{'}) u_i\big(f(\theta_i^{'},\theta_{-i}^{'}),(\theta_i,\theta_{-i})\big).$$

Definition 4.2. The SCF f satisfies *weak robust monotonicity (weak RM)* if every unacceptable deception β is weakly refutable.

Here is the main result of this section (the proof is in the Appendix):

Theorem 4.3. *If the SCF f is RoRat-implementable, then f satisfies weak RM.*

BM (2011) identify *strict* robust monotonicity as a necessary condition for Rat-implementation, and hence, for RoEq-implementation (because the latter implies the former). We present an equivalent definition below.

Definition 4.4. We say that an unacceptable deception β is *strictly refutable* if there exist $i \in I$, $\theta_i \in \Theta_i$, and $\theta_i' \in \beta_i(\theta_i)$ satisfying $\theta_i' \not\sim_i^f \theta_i$ such that for all $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ satisfying $\psi_i(\theta_{-i}, \theta_{-i}') > 0 \Rightarrow \theta_{-i}' \in \beta_{-i}(\theta_{-i})$, there exists $y \in \bigcap_{\tilde{\theta}_i \in \Theta_i} Y_i[\tilde{\theta}_i]$ such that

$$\sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(y(\theta_{-i}'), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(f(\theta_i', \theta_{-i}'), (\theta_i, \theta_{-i})\big).$$

Definition 4.5. The SCF f satisfies *strict robust monotonicity (strict RM)* if every unacceptable deception β is strictly refutable.

Strict RM implies weak RM since the former imposes a stronger refutability requirement on every unacceptable deception, i.e., if an unacceptable deception β is strictly refutable, then it is weakly refutable. This is because strict refutability requires us to find a y in $\bigcap_{\tilde{\theta}_i \in \Theta_i} Y_i[\tilde{\theta}_i]$ whereas for weak refutability, we are allowed to find a y in $Y_i[\tilde{\theta}_i]$ that depends on $\tilde{\theta}_i$. See Section 1.1, where we explain how this difference in the two refutability requirements stems from the difference in the underlying solution concepts, interim equilibrium for RoEq-implementation whereas rationalizability for RoRat-implementation.

BM (2010, Lemmas 4, 5, and 6 and Proposition 4) show that strict RM is necessary for wRat-implementation of responsive SCFs. As wRat-implementation is equivalent to RoRat-implementation (Theorem 3.1), we conclude that, for responsive SCFs, strict RM is a necessary condition for RoRat-implementation. This conclusion is consistent with Theorem 4.3 because, for responsive SCFs, weak RM is equivalent to strict RM, as noted in the next lemma (the proof is in the Appendix).

Lemma 4.6. *Suppose the SCF f is responsive. Then f satisfies strict RM if and only if f satisfies weak RM.*

However, there are non-responsive SCFs that satisfy weak RM but not strict RM, as shown in Example 6.1.

5 Sufficiency for RoRat-Implementation

In this section, we show that weak RM is sufficient for RoRat-implementation under a mild additional assumption: conditional no total indifference (as discussed below, our definition is weaker than the one appearing in BM, 2011).

For each $i \in I$ and $\theta_i \in \Theta_i$, define

$$Y_i^w[\theta_i] \equiv \left\{ y : \Theta_{-i} \to \Delta(A) : \forall \theta_{-i},\ u_i\big(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\big) \geq u_i\big(y(\theta_{-i}), (\theta_i, \theta_{-i})\big) \right\}.$$

Thus, $Y_i^w[\theta_i]$ is the collection of all mappings $y : \Theta_{-i} \to \Delta(A)$ such that for every $\theta_{-i} \in \Theta_{-i}$, the lottery $y(\theta_{-i})$ is weakly worse than $f(\theta_i, \theta_{-i})$ for individual i in state (θ_i, θ_{-i}). Notice that $Y_i[\theta_i]$ (recall the definition from Section 4) is a subset of $Y_i^w[\theta_i]$.

Definition 5.1. The SCF f satisfies *conditional no total indifference* (conditional NTI) if, for all $i \in I,\ \theta_i \in \Theta_i$, and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$, there exist $y, y' \in Y_i^w[\theta_i]$ such that

$$\sum_{\theta_{-i}, \theta'_{-i}} \psi_i(\theta_{-i}, \theta'_{-i}) u_i\big(y(\theta'_{-i}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta'_{-i}} \psi_i(\theta_{-i}, \theta'_{-i}) u_i\big(y'(\theta'_{-i}), (\theta_i, \theta_{-i})\big).$$

Remark: BM (2011) also define a "conditional no total indifference" condition which is stronger than our definition. They require the existence of the said y and y' in the set $\bigcap_{\tilde{\theta}_i \in \Theta_i} Y_i^w[\tilde{\theta}_i]$ whereas we only require the existence of y and y' in the set $Y_i^w[\theta_i]$.

In the sufficiency result, we focus on a countable subset of $Y_i^w[\theta_i]$, as defined next. Recall that $\Delta^*(A)$ is a countable dense subset of $\Delta(A)$. For each i and θ_i, define

$$Y_i^*[\theta_i] \equiv \left\{ y : \Theta_{-i} \to \Delta(A) : \begin{array}{ll} \forall \theta_{-i}, & \\ \text{(i)} & y(\theta_{-i}) \in \Delta^*(A) \bigcup_{\theta'_i \in \Theta_i} \{f(\theta'_i, \theta_{-i})\} \text{ and} \\ \text{(ii)} & u_i\big(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\big) \geq u_i\big(y(\theta_{-i}), (\theta_i, \theta_{-i})\big) \end{array} \right\}$$

Note that $Y_i^*[\theta_i] \subseteq Y_i^w[\theta_i]$. Since Θ_{-i} is finite and $\Delta^*(A)$ is countable, $Y_i^*[\theta_i]$ is also countable. Thus, we denote $Y_i^*[\theta_i]$ by $\{y_i^0[\theta_i], y_i^1[\theta_i], \ldots, y_i^k[\theta_i], \ldots\}$. For each $i \in I$ and $\theta_i \in \Theta_i$, we then define $y_i^{\theta_i} : \Theta_{-i} \to \Delta A$ such that

$$y_i^{\theta_i}(\theta_{-i}) = (1 - \delta) \sum_{k=0}^{\infty} \delta^k y_i^k[\theta_i](\theta_{-i}), \forall \theta_{-i},$$

where $\delta \in (0, 1)$.

Similarly, since A is countable, we denote it by $\{a_0, a_1, \ldots, a_k, \ldots\}$. Then, we define

$$\bar{\alpha} = (1-\eta)\sum_{k=0}^{\infty} \eta^k a_k,$$

where $\eta \in (0,1)$.

The following lemma notes two important consequences of conditional NTI (the proof is in the Appendix).

Lemma 5.2. *If the SCF f satisfies conditional NTI, then the following statements are true:*

(a) For all $i \in I$, $\theta_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$, there exists $y \in Y_i^*[\theta_i]$ such that

$$\sum_{\theta_{-i}, \theta'_{-i}} \psi_i(\theta_{-i}, \theta'_{-i}) u_i\big(y(\theta'_{-i}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta'_{-i}} \psi_i(\theta_{-i}, \theta'_{-i}) u_i\big(y_i^{\theta_i}(\theta'_{-i}), (\theta_i, \theta_{-i})\big).$$

(b) For all $i \in I$, $\theta_i \in \Theta_i$ and $z_i^1 \in \Delta(\Theta_{-i})$, there exists $a \in A$ such that

$$\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(a, (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(\bar{\alpha}, (\theta_i, \theta_{-i})\big).$$

We need one more result before presenting our main sufficiency result for this section.

Definition 5.3. The SCF f satisfies *ex post incentive compatibility* (EPIC) if, for all $i \in I$, $\theta_i, \theta'_i \in \Theta_i$, and $\theta_{-i} \in \Theta_{-i}$,

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

The SCF f satisfies *semi-strict ex post incentive compatible* (semi-strict EPIC) if the above inequality becomes strict whenever $\theta_i \not\sim_i^f \theta'_i$.

We now show that weak RM implies semi-strict EPIC (the proof is in the Appendix):

Lemma 5.4. *If the SCF f satisfies weak RM, then it satisfies semi-strict EPIC.*[5]

[5]BM (2010, Lemma 6) show that if f is wRat-implementable, then it satisfies semi-strict EPIC. It follows from their result and our Theorem 3.1 that semi-strict EPIC is a necessary condition for RoRat-implementation. The above lemma does not immediately follow from BM's result because weak RM is a necessary condition for wRat-implementation. Moreover, due to this lemma, we do not have to add semi-strict EPIC as an additional condition in our sufficiency result. BM (2011, Lemma 1) show that "robust monotonicity" implies semi-strict EPIC. Robust monotonicity is a slightly weaker version of strict RM – the only difference is that we need to replace "$y \in \bigcap_{\tilde{\theta}_i} Y_i[\tilde{\theta}_i]$" with "$y \in \bigcap_{\tilde{\theta}_i} Y_i^w[\tilde{\theta}_i]$" in the definition of strict refutability. Strictly speaking, weak RM and robust monotonicity are not comparable.

For the sufficiency result, we propose the following mechanism $\Gamma = ((M_i)_{i\in I}, g)$: For each individual i, pick any one payoff type from Θ_i. We denote this payoff type as θ_i^*. Each individual i sends a message $m_i = (m_i^1, m_i^2, m_i^3, m_i^4)$, where $m_i^1 = (m_i^1[j])_{j\in I}$ such that $m_i^1[j] \in \Theta_j$ for all $j \in I$, $m_i^2 \in \mathbb{N}, m_i^3 = (m_i^3[\theta_i])_{\theta_i\in\Theta_i}$ such that $m_i^3[\theta_i] \in Y_i^*[\theta_i]$ for all $\theta_i \in \Theta_i$, and $m_i^4 \in A$. Note that each M_i is countable. The outcome function $g : M \to \Delta(A)$ is defined as follows: For each $m \in M$,

Rule 1: $m_i^2 = 1$ for all $i \in I \Rightarrow g(m) = f(m_1^1[1], m_2^1[2], \ldots, m_n^1[n])$.

Rule 2: If there exists $i \in I$ such that $m_i^2 > 1$ but $m_j^2 = 1$ for all $j \in I\backslash\{i\}$, then one of the following sub-rules apply:

Rule 2-1: If there exists $\theta_i \in \Theta_i$ such that $m_j^1[i] = \theta_i$ for all $j \in I\backslash\{i\}$, then

$$g(m) = \begin{cases} m_i^3[\theta_i]\big((m_j^1[j])_{j\neq i}\big) & \text{with probability } m_i^2/(m_i^2+1), \\ y_i^{\theta_i}\big((m_j^1[j])_{j\neq i}\big) & \text{with probability } 1/(m_i^2+1). \end{cases}$$

Rule 2-2: If $m_{j'}^1[i] \neq m_k^1[i]$ for some $j', k \in I\backslash\{i\}$, then

$$g(m) = \begin{cases} m_i^3[\theta_i^*]\big((m_j^1[j])_{j\neq i}\big) & \text{with probability } m_i^2/(m_i^2+1), \\ y_i^{\theta_i^*}\big((m_j^1[j])_{j\neq i}\big) & \text{with probability } 1/(m_i^2+1). \end{cases}$$

Rule 3: In all other cases:

$$g(m) = \begin{cases} m_1^4 & \text{with probability } m_1^2/(1+m_1^2)n, \\ m_2^4 & \text{with probability } m_2^2/(1+m_2^2)n, \\ \vdots & \vdots \\ m_n^4 & \text{with probability } m_n^2/(1+m_n^2)n, \\ \bar{\alpha} & \text{with remaining probability.} \end{cases}$$

Here is our sufficiency result for RoRat-implementation (the proof is in the Appendix):

Theorem 5.5. *If the SCF f satisfies weak RM and conditional NTI, then it is RoRat-implementable.*

Remark 5.6. Although the mechanism constructed to prove the above sufficiency result does share aspects with standard canonical constructions, it is worth pointing out one of its distinctive features (compare, for instance, to the mechanism in BM, 2011): Each player reports a payoff state, i.e., not just her own but also everyone else's payoff type. To see the

importance of this, consider two types t_i and t_i' of agent i with distinct payoff types, say θ_i and θ_i', respectively. Moreover, suppose that both t_i and t_i' agree on the payoff types of everyone else, say θ_{-i}. Then, from the perspective of t_i, the true payoff state is (θ_i, θ_{-i}) whereas from the perspective of t_i', the true payoff state is (θ_i', θ_{-i}). Since their truths are different, these two types cannot both be correct if they believe that everyone else is reporting the *payoff state* truthfully. While this is problematic for truthful behavior to form an equilibrium, it does not cause any issues for truthful behavior to be rationalizable because rationalizability does not require the two types to hold common beliefs about the other agents' behavior. This is precisely the kind of flexibility that is needed in order to RoRat-implement an SCF that cannot be RoEq-implemented, as illustrated in Section 1.1.

6 RoRat-Implementation versus RoEq-Implementation

Recall the discussion in the Introduction where we pointed out that RoRat-implementation is weaker than RoEq-implementation and asked the question whether the former could be strictly weaker than the latter. In Section 6.1, we answer that question in the affirmative by providing an example of an SCF that is RoRat-implementable but *not* RoEq-implementable.

As discussed earlier, the uniqueness requirement in RoRat-implementation is the same as that in RoEq-implementation. Thus, the explanation for the gap between the two implementation notions is that the nonemptiness requirement in RoEq-implementation is strictly stronger than that in RoRat-implementation. Any mechanism, in particular the canonical mechanism in the proof of Theorem 5.5, that RoRat-implements an SCF which is not RoEq-implementable must fail the nonemptiness requirement for RoEq-implementation. That is, there must exist *some* type space in which the set of interim equilibria of the mechanism is empty. However, this does not preclude the possibility that the mechanism has nonempty interim equilibria on several other type spaces. Indeed, in Section 6.2, we establish that our canonical mechanism that RoRat-implements the desired SCF has nonempty interim equilibria on type spaces that are typically found in the applied literature.

There are two notable cases when RoRat-implementation is equivalent to RoEq-implementation: First, when we restrict the designer to finite mechanisms and second, when the environment is one of private values. We discuss these cases in Section 6.3 and Section 6.4, respectively.

6.1 An Example

We now present an example with an SCF that is RoRat-implementable but not RoEq-implementable. We do so by exploiting the gap between strict RM and weak RM for non-

responsive SCFs. (Strict RM and weak RM are equivalent for responsive SCFs, as shown in Lemma 4.6.) To elaborate, BM (2011) show that strict RM is a necessary condition for Rat-implementation of any SCF. The non-responsive SCF in the example below fails to satisfy strict RM. Thus, the SCF is not Rat-implementable, and hence not RoEq-implementable. The SCF, however, satisfies weak RM and conditional NTI. Hence, the SCF is RoRat-implementable.

Example 6.1. There are two players $i \in \{1, 2\}$. Player 1 has three payoff types: $\Theta_1 = \{\theta_1, \theta_1', \theta_1''\}$ and player 2 has two payoff types: $\Theta_2 = \{\theta_2, \theta_2'\}$. There are six pure alternatives: $A = \{a, b, c, d, z, z'\}$. The following tables list the payoffs of the two players:

a	θ_2	θ_2'
θ_1	$4, 4$	$4, 0$
θ_1'	$0, 0$	$4, 1$
θ_1''	$1, 1$	$4, 0$

b	θ_2	θ_2'
θ_1	$0, 0$	$3, 3$
θ_1'	$1, 1$	$2, 0$
θ_1''	$0, 0$	$2, 1$

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

d	θ_2	θ_2'
θ_1	$3, 4$	$2, 0$
θ_1'	$0, 0$	$3, 3$
θ_1''	$0, 0$	$3, 3$

z	θ_2	θ_2'
θ_1	$4, 1$	$2, 0$
θ_1'	$2, 2$	$5, 0$
θ_1''	$2, 2$	$2, 0$

z'	θ_2	θ_2'
θ_1	$4, 0$	$4, 1$
θ_1'	$2, 0$	$2, 2$
θ_1''	$2, 0$	$5, 0$

The SCF f selects the alternative which maximizes the aggregate payoff in each payoff state.

f	θ_2	θ_2'
θ_1	a	b
θ_1'	c	d
θ_1''	c	d

We first show that f fails strict RM.

Claim 6.2. *The SCF f violates strict RM.*

Proof. Consider the unacceptable deception β such that

$$\beta_1(\theta_1) = \{\theta_1, \theta_1'\}, \quad \beta_1(\theta_1') = \{\theta_1'\}, \quad \beta_1(\theta_1'') = \{\theta_1''\},$$

and

$$\beta_2(\theta_2) = \{\theta_2, \theta_2'\}, \quad \beta_2(\theta_2') = \{\theta_2'\}.$$

Given this deception, there are exactly two tuples (i, θ_i, θ_i') such that $\theta_i' \in \beta_i(\theta_i)$ and $\theta_i' \not\sim_i^f \theta_i$: $(1, \theta_1, \theta_1')$ and $(2, \theta_2, \theta_2')$.

First, consider $(2, \theta_2, \theta_2')$. Fix the degenerate belief $\psi_2 \in \Delta(\Theta_1 \times \Theta_1)$ such that $\psi_2(\theta_1, \theta_1') = 1$. Then, there does not exist any $y \in \bigcap_{\tilde{\theta}_2 \in \Theta_2} Y_2[\tilde{\theta}_2]$ such that

$$u_2\big(y(\theta_1'), (\theta_1, \theta_2)\big) > u_2\big(f(\theta_1', \theta_2'), (\theta_1, \theta_2)\big),$$

because $f(\theta_1', \theta_2') = d$ is one of the best alternatives for player 2 in the payoff state (θ_1, θ_2).

Second, consider $(1, \theta_1, \theta_1')$. Fix the degenerate belief ψ_1 such that $\psi_1(\theta_2, \theta_2') = 1$. If there exists $y \in \bigcap_{\tilde{\theta}_1 \in \Theta_1} Y_1[\tilde{\theta}_1]$, then $y(\theta_2')$ must satisfy the following equations

$$\begin{aligned} u_1\big(f(\theta_1', \theta_2'), (\theta_1', \theta_2')\big) &\geq u_1\big(y(\theta_2'), (\theta_1', \theta_2')\big) \\ u_1\big(f(\theta_1'', \theta_2'), (\theta_1'', \theta_2')\big) &\geq u_1\big(y(\theta_2'), (\theta_1'', \theta_2')\big). \end{aligned}$$

These two inequalities imply that

$$2y(\theta_2')[z] + y(\theta_2')[a] \leq y(\theta_2')[z'] + y(\theta_2')[b] \quad \text{and} \quad 2y(\theta_2')[z'] + y(\theta_2')[a] \leq y(\theta_2')[z] + y(\theta_2')[b],$$

where $y(\theta_2')[x]$ is the probability of alternative x in the lottery $y(\theta_2')$. Summing these two inequalities, we obtain $y(\theta_2')[z] + y(\theta_2')[z'] + 2y(\theta_2')[a] \leq 2y(\theta_2')[b]$. In order to satisfy strict RM, we must satisfy the following inequality:

$$u_1\big(y(\theta_2'), (\theta_1, \theta_2)\big) > u_1\big(f(\theta_1', \theta_2'), (\theta_1, \theta_2)\big).$$

The above inequality is translated into $y(\theta_2')[z] + y(\theta_2')[z'] + y(\theta_2')[a] > 3y(\theta_2')[b] + 3y(\theta_2')[c]$. We then claim that this inequality is impossible to satisfy. Plugging $y(\theta_2')[z] + y(\theta_2')[z'] + 2y(\theta_2')[a] \leq 2y(\theta_2')[b]$ into $y(\theta_2')[z] + y(\theta_2')[z'] + y(\theta_2')[a] > 3y(\theta_2')[b] + 3y(\theta_2')[c]$, we obtain

$$-y(\theta_2')[a] > y(\theta_2')[b] + 3y(\theta_2')[c].$$

However, this inequality is impossible because $y(\theta_2')[a], y(\theta_2')[b]$, and $y(\theta_2')[c]$ all are nonnegative. We therefore conclude that the SCF f does not satisfy strict RM. □

Next we argue that f satisfies weak RM.

Claim 6.3. *The SCF f satisfies weak RM.*

Proof. First, we consider any unacceptable deception β such that either $\theta_1' \in \beta_1(\theta_1)$ or $\theta_1'' \in \beta_1(\theta_1)$. Pick any belief $\psi_1 \in \Delta(\Theta_2 \times \Theta_2)$. Then

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(f(\theta_1', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) \;=\; \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(f(\theta_1'', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big)$$

$$= \quad 3\psi_1(\theta_2', \theta_2) + 3\psi_1(\theta_2, \theta_2') + 2\psi_1(\theta_2', \theta_2'),$$

where the first equality follows from the fact that f is non-responsive to θ_1' and θ_1''.

In what follows, we consider each possible case of $\tilde{\theta}_1 \in \{\theta_1, \theta_1', \theta_1''\}$.

Case 1: $\tilde{\theta}_1 = \theta_1$.

Define $y : \Theta_2 \to \Delta(A)$ to be such that $y(\theta_2) = a$ and $y(\theta_2') = \frac{2}{3}z + \frac{1}{3}z'$. It is straightforward to confirm that $y \in Y_1[\theta_1]$. Moreover,

$$\begin{aligned}
\sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y(\tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) &= 4\psi_1(\theta_2, \theta_2) + 4\psi_1(\theta_2', \theta_2) + 4\psi_1(\theta_2, \theta_2') + \frac{8}{3}\psi_1(\theta_2', \theta_2') \\
&> 3\psi_1(\theta_2', \theta_2) + 3\psi_1(\theta_2, \theta_2') + 2\psi_1(\theta_2', \theta_2') \\
&= \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(f(\theta_1', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) \\
&= \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(f(\theta_1'', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big).
\end{aligned}$$

Case 2: $\tilde{\theta}_1 = \theta_1'$.

Define $y : \Theta_2 \to \Delta(A)$ to be such that $y(\theta_2) = a$ and $y(\theta_2') = z'$. It is straightforward to confirm that $y \in Y_1[\theta_1']$. Moreover,

$$\begin{aligned}
\sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y(\tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) &= 4\psi_1(\theta_2, \theta_2) + 4\psi_1(\theta_2', \theta_2) + 4\psi_1(\theta_2, \theta_2') + 4\psi_1(\theta_2', \theta_2') \\
&> 3\psi_1(\theta_2', \theta_2) + 3\psi_1(\theta_2, \theta_2') + 2\psi_1(\theta_2', \theta_2') \\
&= \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(f(\theta_1', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) \\
&= \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(f(\theta_1'', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big).
\end{aligned}$$

Case 3: $\tilde{\theta}_1 = \theta_1''$.

Define $y : \Theta_2 \to \Delta(A)$ to be such that $y(\theta_2) = a$ and $y(\theta_2') = \frac{1}{5}c + \frac{4}{5}z$. It is straightforward to confirm that $y \in Y_1[\theta_1'']$. Moreover,

$$\begin{aligned}
\sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y(\tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) &= 4\psi_1(\theta_2, \theta_2) + 4\psi_1(\theta_2', \theta_2) + \frac{16}{5}\psi_1(\theta_2, \theta_2') + \frac{11}{5}\psi_1(\theta_2', \theta_2') \\
&> 3\psi_1(\theta_2', \theta_2) + 3\psi_1(\theta_2, \theta_2') + 2\psi_1(\theta_2', \theta_2')
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(f(\theta_1', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) \\
&= \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(f(\theta_1'', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big).
\end{aligned}
$$

It follows that any unacceptable deception β satisfying $\theta_1' \in \beta_1(\theta_1)$ is weakly refutable using the tuple $(1, \theta_1, \theta_1')$ whereas any unacceptable deception β satisfying $\theta_1'' \in \beta_1(\theta_1)$ is weakly refutable using the tuple $(1, \theta_1, \theta_1'')$.

Second, we consider any unacceptable deception β such that $\theta_2' \in \beta_2(\theta_2)$ and $\beta_1(\theta_1) = \{\theta_1\}$. Pick any belief $\psi_2 \in \Delta(\Theta_1 \times \Theta_1)$ such that $\psi_2(\tilde{\theta}_1, \tilde{\theta}_1') > 0 \Rightarrow \tilde{\theta}_1' \in \beta_1(\tilde{\theta}_1)$. Then we have $\psi_2(\theta_1, \theta_1') = \psi_2(\theta_1, \theta_1'') = 0$. Therefore,

$$
\sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2\big(f(\tilde{\theta}_1', \theta_2'), (\tilde{\theta}_1, \theta_2)\big) = \psi_2(\theta_1', \theta_1).
$$

Define $y : \Theta_1 \to \Delta(A)$ to be such that $y(\theta_1) = y(\theta_1') = y(\theta_1'') = z$. It is straightforward to confirm that $y \in Y_2[\theta_2] \cap Y_2[\theta_2']$. Moreover, since $\psi_2(\theta_1, \theta_1') = \psi_2(\theta_1, \theta_1'') = 0$, we have

$$
\begin{aligned}
&\sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2\big(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2)\big) \\
= \;& \psi_2(\theta_1, \theta_1) + 2\big(\psi_2(\theta_1', \theta_1) + \psi_2(\theta_1', \theta_1') + \psi_2(\theta_1', \theta_1'')\big) + 2\big(\psi_2(\theta_1'', \theta_1) + \psi_2(\theta_1'', \theta_1') + \psi_2(\theta_1'', \theta_1'')\big) \\
> \;& \psi_2(\theta_1', \theta_1) = \sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2\big(f(\tilde{\theta}_1', \theta_2'), (\tilde{\theta}_1, \theta_2)\big).
\end{aligned}
$$

It follows that any unacceptable deception β such that $\theta_2' \in \beta_2(\theta_2)$ and $\beta_1(\theta_1) = \{\theta_1\}$ is weakly refutable using the tuple $(2, \theta_2, \theta_2')$.

Third, we consider any unacceptable deception β such that $\theta_2 \in \beta_2(\theta_2')$ and $\beta_1(\theta_1) = \{\theta_1\}$. Pick any belief $\psi_2 \in \Delta(\Theta_1 \times \Theta_1)$ such that $\psi_2(\tilde{\theta}_1, \tilde{\theta}_1') > 0 \Rightarrow \tilde{\theta}_1' \in \beta_1(\tilde{\theta}_1)$. Then we have that $\psi_2(\theta_1, \theta_1') = \psi_2(\theta_1, \theta_1'') = 0$. Therefore,

$$
\sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2\big(f(\tilde{\theta}_1', \theta_2), (\tilde{\theta}_1, \theta_2')\big) = \psi_2(\theta_1', \theta_1).
$$

Define $y : \Theta_1 \to \Delta(A)$ to be such that $y(\theta_1) = y(\theta_1') = y(\theta_1'') = \frac{1}{4}b + \frac{3}{4}z'$. It is straightforward to confirm that $y \in Y_2[\theta_2] \cap Y_2[\theta_2']$. Moreover, since $\psi_2(\theta_1, \theta_1') = \psi_2(\theta_1, \theta_1'') = 0$, we

have

$$\begin{aligned}
&\sum_{\tilde{\theta}_1,\tilde{\theta}_1'} \psi_2(\tilde{\theta}_1,\tilde{\theta}_1')u_2\big(y(\tilde{\theta}_1'),(\tilde{\theta}_1,\theta_2')\big) \\
= \;& \frac{3}{2}\psi_2(\theta_1,\theta_1)+\frac{3}{2}\big(\psi_2(\theta_1',\theta_1)+\psi_2(\theta_1',\theta_1')+\psi_2(\theta_1',\theta_1'')\big)+\frac{1}{4}\big(\psi_2(\theta_1'',\theta_1)+\psi_2(\theta_1'',\theta_1')+\psi_2(\theta_1'',\theta_1'')\big) \\
> \;& \psi_2(\theta_1',\theta_1)=\sum_{\tilde{\theta}_1,\tilde{\theta}_1'} \psi_2(\tilde{\theta}_1,\tilde{\theta}_1')u_2\big(f(\tilde{\theta}_1',\theta_2),(\tilde{\theta}_1,\theta_2')\big).
\end{aligned}$$

It follows that any unacceptable deception β such that $\theta_2 \in \beta_2(\theta_2')$ and $\beta_1(\theta_1)=\{\theta_1\}$ is weakly refutable using the tuple $(2,\theta_2',\theta_2)$.

Fourth, we consider any unacceptable deception such that $\beta_1(\theta_1)=\{\theta_1\}$, $\beta_2(\theta_2)=\{\theta_2\}$, and $\beta_2(\theta_2')=\{\theta_2'\}$. Such a deception involves either $\theta_1 \in \beta_1(\theta_1')$ or $\theta_1 \in \beta_1(\theta_1'')$. Then the fact that f satisfies semi-strict EPIC implies that β is weakly refutable. We show this formally for the case when $\theta_1 \in \beta_1(\theta_1')$ as the argument for the case when $\theta_1 \in \beta_1(\theta_1'')$ is similar. So suppose $\theta_1 \in \beta_1(\theta_1')$. Pick any belief $\psi_1 \in \Delta(\Theta_2 \times \Theta_2)$ such that $\psi_1(\tilde{\theta}_2,\tilde{\theta}_2')>0 \Rightarrow \tilde{\theta}_2' \in \beta_2(\tilde{\theta}_2)$. Then we have that $\psi_1(\theta_2,\theta_2')=\psi_1(\theta_2',\theta_2)=0$. Therefore,

$$\sum_{\tilde{\theta}_2,\tilde{\theta}_2'} \psi_1(\tilde{\theta}_2,\tilde{\theta}_2')u_1\big(f(\theta_1,\tilde{\theta}_2'),(\theta_1',\tilde{\theta}_2)\big)=2\psi_1(\theta_2',\theta_2')$$

Define $y:\Theta_2 \to \Delta(A)$ to be such that $y(\theta_2)=f(\theta_1',\theta_2)=c$ and $y(\theta_2')=f(\theta_1',\theta_2')=d$. It is straightforward to confirm that $y \in Y_1[\theta_1]\cap Y_1[\theta_1']\cap Y_1[\theta_1'']$. Moreover, since $\psi_1(\theta_2,\theta_2')=\psi_1(\theta_2',\theta_2)=0$,

$$\sum_{\tilde{\theta}_2,\tilde{\theta}_2'} \psi_1(\tilde{\theta}_2,\tilde{\theta}_2')u_1\big(y(\tilde{\theta}_2'),(\theta_1',\tilde{\theta}_2)\big)=3\psi_1(\theta_2,\theta_2)+3\psi_1(\theta_2',\theta_2')>\sum_{\tilde{\theta}_2,\tilde{\theta}_2'} \psi_1(\tilde{\theta}_2,\tilde{\theta}_2')u_1\big(f(\theta_1,\tilde{\theta}_2'),(\theta_1',\tilde{\theta}_2)\big).$$

It follows that the deception β is weakly refutable using the tuple $(1,\theta_1',\theta_1)$.

We thus conclude that every unacceptable deception is weakly refutable, and hence f satisfies weak RM. □

We now check that the SCF f satisfies conditional NTI.

Claim 6.4. *The SCF f satisfies conditional NTI.*

Proof. First, we consider player 1 of payoff type θ_1. Let $y:\Theta_2 \to \Delta(A)$ be such that $y(\theta_2)=a$ and $y(\theta_2')=z$. Also, let $y':\Theta_2 \to \Delta(A)$ be such that $y'(\theta_2)=b$ and $y'(\theta_2')=d$. It

is straightforward to confirm that $y, y^{'} \in Y_1^w[\theta_1]$. Now,

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2^{'}} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2^{'}) u_1\big(y(\tilde{\theta}_2^{'}), (\theta_1, \tilde{\theta}_2)\big) = 4\psi_1(\theta_2, \theta_2) + 4\psi_1(\theta_2, \theta_2^{'}) + 4\psi_1(\theta_2^{'}, \theta_2) + 2\psi_1(\theta_2^{'}, \theta_2^{'}).$$

whereas

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2^{'}} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2^{'}) u_1\big(y^{'}(\tilde{\theta}_2^{'}), (\theta_1, \tilde{\theta}_2)\big) = 3\psi_1(\theta_2, \theta_2^{'}) + 3\psi_1(\theta_2^{'}, \theta_2) + 2\psi_1(\theta_2^{'}, \theta_2^{'}).$$

We therefore have that

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2^{'}} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2^{'}) u_1\big(y(\tilde{\theta}_2^{'}), (\theta_1, \tilde{\theta}_2)\big) = \sum_{\tilde{\theta}_2, \tilde{\theta}_2^{'}} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2^{'}) u_1\big(y^{'}(\tilde{\theta}_2^{'}), (\theta_1, \tilde{\theta}_2)\big) \Leftrightarrow \psi_1(\theta_2^{'}, \theta_2^{'}) = 1.$$

Thus, for all $\psi_1 \in \Delta(\Theta_2 \times \Theta_2)$ such that $\psi_1(\theta_2^{'}, \theta_2^{'}) < 1$, we have found $y, y^{'} \in Y_1^w[\theta_1]$ that satisfy the requirement for conditional NTI. If ψ_1 is such that $\psi_1(\theta_2^{'}, \theta_2^{'}) = 1$, then we define $y : \Theta_2 \to \Delta(A)$ such that $y(\theta_2) = y(\theta_2^{'}) = b$ and $y^{'} : \Theta_2 \to \Delta(A)$ such that $y^{'}(\theta_2) = y^{'}(\theta_2^{'}) = d$. It is straightforward to confirm that $y, y^{'} \in Y_1^w[\theta_1]$. Since $\psi_1(\theta_2^{'}, \theta_2^{'}) = 1$, $u_1(y(\theta_2^{'}), (\theta_1, \theta_2^{'})) = u_1(b, (\theta_1, \theta_2^{'})) = 3$ and $u_1(y^{'}(\theta_2^{'}), (\theta_1, \theta_2^{'})) = u_1(d, (\theta_1, \theta_2^{'})) = 2$, we obtain

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2^{'}} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2^{'}) u_1\big(y(\tilde{\theta}_2^{'}), (\theta_1, \tilde{\theta}_2)\big) > \sum_{\tilde{\theta}_2, \tilde{\theta}_2^{'}} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2^{'}) u_1\big(y^{'}(\tilde{\theta}_2^{'}), (\theta_1, \tilde{\theta}_2)\big).$$

Thus, if ψ_1 is such that $\psi_1(\theta_2^{'}, \theta_2^{'}) = 1$, then too we satisfy the requirement for conditional NTI.

Second, we consider player 1 of payoff type $\theta_1^{'}$. Then we define $y : \Theta_2 \to \Delta(A)$ such that $y(\theta_2) = y(\theta_2^{'}) = c$ and $y^{'} : \Theta_2 \to \Delta(A)$ such that $y^{'}(\theta_2) = y^{'}(\theta_2^{'}) = b$. It is straightforward to confirm that $y, y^{'} \in Y_1^w[\theta_1^{'}]$. Fix $\psi_1 \in \Delta(\Theta_2 \times \Theta_2)$. Now,

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2^{'}} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2^{'}) u_1\big(y(\tilde{\theta}_2^{'}), (\theta_1^{'}, \tilde{\theta}_2)\big) = 3\psi_1(\theta_2, \theta_2) + 3\psi_1(\theta_2, \theta_2^{'}) + 3\psi_1(\theta_2^{'}, \theta_2) + 3\psi_1(\theta_2^{'}, \theta_2^{'})$$

whereas

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2^{'}} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2^{'}) u_1\big(y^{'}(\tilde{\theta}_2^{'}), (\theta_1^{'}, \tilde{\theta}_2)\big) = \psi_1(\theta_2, \theta_2) + \psi_1(\theta_2, \theta_2^{'}) + 2\psi_1(\theta_2^{'}, \theta_2) + 2\psi_1(\theta_2^{'}, \theta_2^{'}).$$

This implies that for any $\psi_1 \in \Delta(\Theta_2 \times \Theta_2)$,

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y(\tilde{\theta}_2'), (\theta_1', \tilde{\theta}_2)\big) > \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y'(\tilde{\theta}_2'), (\theta_1', \tilde{\theta}_2)\big).$$

Thus, we satisfy the requirement for conditional NTI.

Third, we consider player 1 of payoff type θ_1''. Once again, we define $y : \Theta_2 \to \Delta(A)$ such that $y(\theta_2) = y(\theta_2') = c$ and $y' : \Theta_2 \to \Delta(A)$ such that $y'(\theta_2) = y'(\theta_2') = b$. It is straightforward to confirm that $y, y' \in Y_1^w[\theta_1'']$. Fix $\psi_1 \in \Delta(\Theta_2 \times \Theta_2)$. Now

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y(\tilde{\theta}_2'), (\theta_1'', \tilde{\theta}_2)\big) = 3\psi_1(\theta_2, \theta_2) + 3\psi_1(\theta_2, \theta_2') + 3\psi_1(\theta_2', \theta_2) + 3\psi_1(\theta_2', \theta_2')$$

whereas

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y'(\tilde{\theta}_2'), (\theta_1'', \tilde{\theta}_2)\big) = 2\psi_1(\theta_2', \theta_2) + 2\psi_1(\theta_2', \theta_2').$$

This implies that for any $\psi_1 \in \Delta(\Theta_2 \times \Theta_2)$,

$$\sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y(\tilde{\theta}_2'), (\theta_1'', \tilde{\theta}_2)\big) > \sum_{\tilde{\theta}_2, \tilde{\theta}_2'} \psi_1(\tilde{\theta}_2, \tilde{\theta}_2') u_1\big(y'(\tilde{\theta}_2'), (\theta_1'', \tilde{\theta}_2)\big).$$

Thus, we satisfy the requirement for conditional NTI.

Fourth, we consider player 2 of payoff type θ_2. Then we define $y : \Theta_1 \to \Delta(A)$ such that $y(\theta_1) = y(\theta_1') = y(\theta_1'') = \frac{1}{2}a + \frac{1}{2}c$ and $y' : \Theta_1 \to \Delta(A)$ such that $y'(\theta_1) = y'(\theta_1') = y'(\theta_1'') = b$. It is straightforward to confirm that $y, y' \in Y_2^w[\theta_2]$. Fix $\psi_2 \in \Delta(\Theta_2 \times \Theta_2)$. Now

$$\begin{aligned} & \sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2)) \\ = \; & 2(\psi_1(\theta_1, \theta_1) + \psi_1(\theta_1, \theta_1') + \psi_1(\theta_1, \theta_1'')) + \frac{3}{2}(\psi_2(\theta_1', \theta_1) + \psi_2(\theta_1', \theta_1') + \psi_2(\theta_1', \theta_1'')) \\ & +2(\psi_2(\theta_1'', \theta_1) + \psi_2(\theta_1'', \theta_1') + \psi_2(\theta_1'', \theta_1'')) \end{aligned}$$

whereas

$$\sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2(y'(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2)) = \psi_2(\theta_1', \theta_1) + \psi_2(\theta_1', \theta_1') + \psi_2(\theta_1', \theta_1'').$$

This implies that for any $\psi_2 \in \Delta(\Theta_1 \times \Theta_1)$,

$$\sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2)) > \sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2(y'(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2)).$$

Thus, we satisfy the requirement for conditional NTI.

Finally, we consider player 2 of payoff type θ_2'. Then we define $y : \Theta_1 \to \Delta(A)$ such that $y(\theta_1) = y(\theta_1') = y(\theta_1'') = \frac{1}{2}b + \frac{1}{2}d$ and $y' : \Theta_1 \to \Delta(A)$ such that $y'(\theta_1) = y'(\theta_1') = y'(\theta_1'') = c$. It is straightforward to confirm that $y, y' \in Y_2^w[\theta_2']$. Fix $\psi_2 \in \Delta(\Theta_1 \times \Theta_1)$. Then

$$\begin{aligned} & \sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2')) \\ = \; & \frac{3}{2}(\psi_1(\theta_1, \theta_1) + \psi_1(\theta_1, \theta_1') + \psi_1(\theta_1, \theta_1'')) + \frac{3}{2}(\psi_2(\theta_1', \theta_1) + \psi_2(\theta_1', \theta_1') + \psi_2(\theta_1', \theta_1'')) \\ & +2(\psi_2(\theta_1'', \theta_1) + \psi_2(\theta_1'', \theta_1') + \psi_2(\theta_1'', \theta_1'')) \end{aligned}$$

whereas

$$\sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2(y'(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2')) = \psi_2(\theta_1, \theta_1) + \psi_2(\theta_1, \theta_1') + \psi_2(\theta_1, \theta_1'').$$

This implies that for any $\psi_2 \in \Delta(\Theta_1 \times \Theta_1)$,

$$\sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2')) > \sum_{\tilde{\theta}_1, \tilde{\theta}_1'} \psi_2(\tilde{\theta}_1, \tilde{\theta}_1') u_2(y'(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2')).$$

Thus, we satisfy the requirement for conditional NTI.

We therefore conclude that f satisfies conditional NTI. □

6.2 A Class of Type Spaces in which Our Canonical Mechanism has Interim Equilibria

Consider the following class of information structures: Each individual i is endowed with a signal function $s_i : \Theta \to \Delta(\Theta)$ such that $s_i(\theta)[\theta] > 0$ for any $\theta \in \Theta$ and, for any $\theta, \theta' \in \Theta$, $s_i(\theta)[\theta'] > 0 \Rightarrow \theta_i' = \theta_i$. The profile of signal functions $(s_1, \ldots, s_n)$ is assumed to be common knowledge. An important example of this class of information structures is the one in which $s_i(\theta)[\theta] = 1$ for all $\theta \in \Theta$ and $i \in I$, so that individuals have complete information about the realized state.

A type space $\mathcal{T} = (T_i, \hat{\theta}_i, \hat{\pi}_i)_{i \in I}$ that corresponds to the above class of information struc-

tures is such that (i) $T_i = \{t_i^\theta : \theta \in \Theta\}$ for all $i \in I$ and (ii) $\hat{\theta}_i(t_i^\theta) = \theta_i$, where $\theta = (\theta_i, \theta_{-i})$, and for all $\theta, \theta' \in \Theta$ and $t_{-i}^{\theta'} \in T_{-i}$,

$$\hat{\pi}_i(t_i^\theta)[t_{-i}^{\theta'}] = s_i(\theta)[\theta'],$$

where $t_{-i}^{\theta'} \equiv (t_j^{\theta'})_{j \in I \setminus \{i\}}$. Thus, $\hat{\pi}_i(t_i^\theta)$ is the belief of type t_i^θ such that if $t_{-i} \in T_{-i}$ includes types $t_j^{\tilde{\theta}}$ and $t_{j'}^{\tilde{\theta}'}$ of two distinct individuals $j, j' \in I \setminus \{i\}$ such that $\tilde{\theta} \neq \tilde{\theta}'$, then we have $\hat{\pi}_i(t_i^\theta)[t_{-i}] = 0$.

Suppose the SCF f satisfies weak RM and conditional NTI, so that it is RoRat-implementable by the canonical mechanism Γ constructed in the proof of Theorem 5.5. Pick any type space $\mathcal{T}$, as defined above. We now show that the canonical mechanism Γ has a (pure) interim equilibrium in $\mathcal{T}$.

For each individual $i \in I$, we pick any (m_i^3, m_i^4), where $m_i^3 = (m_i^3[\theta_i])_{\theta_i \in \Theta_i}$ such that $m_i^3[\theta_i] \in Y_i^*[\theta_i]$ for all $\theta_i \in \Theta_i$ and $m_i^4 \in A$. Now let σ be the strategy-profile such that $\sigma_i(t_i^\theta) = (\theta, 1, m_i^3, m_i^4)$ for all $t_i^\theta \in T_i$ and $i \in I$. We argue that σ is an interim equilibrium of the game $(\mathcal{T}, \Gamma)$.

Pick individual i of type t_i^θ. If everyone plays the game $(\Gamma, \mathcal{T})$ according to the strategy profile σ, then the outcome is given by Rule 1 and type t_i^θ of individual i expects a payoff of $\sum_{t_{-i}^{\theta'}} \hat{\pi}_i(t_i^\theta)[t_{-i}^{\theta'}] u_i\big(f(\theta_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big)$.

On the one hand, if type t_i^θ deviates to $\hat{m}_i$ such that $\hat{m}_i^1[i] = \tilde{\theta}_i$ and $\hat{m}_i^2 = 1$, then Rule 1 is still triggered so that she expects the payoff of $\sum_{t_{-i}^{\theta'}} \hat{\pi}_i(t_i^\theta)[t_{-i}^{\theta'}] u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big)$, which is not improving due to semi-strict EPIC. On the other hand, if type t_i^θ deviates to $\hat{m}_i$ such that $\hat{m}_i^2 > 1$, then Rule 2 is triggered so that she expects the payoff of

$$\sum_{t_{-i}^{\theta'}} \hat{\pi}_i(t_i^\theta)[t_{-i}^{\theta'}] \left\{ \left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2} \right) u_i\big(\hat{m}_i^3[\theta_i'](\theta'_{-i}), (\theta_i, \theta'_{-i})\big) + \left(1 - \frac{\hat{m}_i^2}{1+\hat{m}_i^2} \right) u_i\big(y_i^{\theta_i'}(\theta'_{-i}), (\theta_i, \theta'_{-i})\big) \right\}$$

$$= \sum_{t_{-i}^{(\theta_i, \theta'_{-i})}} \hat{\pi}_i(t_i^\theta)[t_{-i}^{(\theta_i, \theta'_{-i})}] \left\{ \left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2} \right) u_i\big(\hat{m}_i^3[\theta_i](\theta'_{-i}), (\theta_i, \theta'_{-i})\big) + \left(1 - \frac{\hat{m}_i^2}{1+\hat{m}_i^2} \right) u_i\big(y_i^{\theta_i}(\theta'_{-i}), (\theta_i, \theta'_{-i})\big) \right\},$$

where the equality follows from the fact that $\hat{\pi}_i(t_i^\theta)[t_{-i}^{\theta'}] = s_i(\theta)[\theta'] > 0 \Rightarrow \theta_i' = \theta_i$. As σ_{-i} dictates that any other agent $j \neq i$ announces $m_j^1[j]$ truthfully and $\hat{m}_i^3[\theta_i]$ is chosen from $Y_i^*[\theta_i]$, type t_i^θ cannot improve her payoff by any such deviation. Hence, the message $\sigma_i(t_i^\theta)$ is a best response of type t_i^θ against σ_{-i}, which completes the argument that σ is an interim equilibrium of the game $(\mathcal{T}, \Gamma)$

6.3 Finite Mechanisms

If we restrict attention to finite mechanisms, then the set of interim equilibria in every countable type space is nonempty.[6] Hence, RoEq-implementation becomes equivalent to RoRat-implementation when the designer is restricted to use finite mechanisms.

BM (2011) show that an additional "robust measurability" condition is necessary for RoEq-implementation using finite mechanisms. It thus follows that robust measurability is also necessary for RoRat-implementation using finite mechanisms. Robust measurability is generally not related to weak RM.[7] It is, therefore, an additional restriction on RoRat-implementation using finite mechanisms. However, for the class of "single crossing aggregator" environments, robust measurability is equivalent to strict RM (see Section 5 in BM (2011) for details). In such environments, any responsive SCF satisfying strict RM can be RoEq-implemented using a direct mechanism in which players report their payoff types (BM, 2009). Recall that strict RM is equivalent to weak RM for responsive SCFs (Lemma 4.6). Thus, it follows that, for responsive SCFs, weak RM by itself characterizes RoRat-implementation using finite mechanisms in single crossing aggregator environments.

In a complete information environment with lotteries and transfers, Chen, Kunimoto, Sun, and Xiong (2020) show that Maskin monotonicity*, a strengthening of Maskin monotonicity, is a necessary and sufficient condition for implementation in rationalizable strategies by a finite mechanism. They also show that Maskin monotonicity* is strictly stronger than Maskin monotonicity, which is a necessary and sufficient condition for Nash implementation by a finite mechanism in the same class of environments with transfers and lotteries (See Chen, Kunimoto, Sun, and Xiong (2019)). Therefore, if we restrict our attention to finite mechanisms in a complete information setup, implementation in rationalizable strategies is more restrictive than Nash implementation. This exhibits a contrast with the equivalence between RoRat-implementation and RoEq-implementation using finite mechanisms.

6.4 Private Values

The environment is one of *private values* if the utility of each player is independent of the other players' payoff types. Thus, player i's utility is a function of the lottery and her own payoff type, i.e., $u_i : \Delta(A) \times \Theta_i \to \Re$.

Semi-strict EPIC is always a necessary condition for RoRat-implementation because weak RM implies semi-strict EPIC (Lemma 5.4). We now argue that it is also sufficient for RoRat-implementation in private values environments.[8]

[6] For a proof of this statement, see Footnote 14 in Bergemann et al. (2017).

[7] We can show this using Examples 1 and 2 in Section 8.3 in BM (2007).

[8] As a corollary, we obtain that weak RM is equivalent to semi-strict EPIC in private values environments.

Under private values, the SCF f satisfies semi-strict EPIC if and only if for all $i \in I$, $\theta_i, \theta_i' \in \Theta_i$, and $\theta_{-i} \in \Theta_{-i}$,

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

with a strict inequality whenever $\theta_i \not\sim_i^f \theta_i'$. Therefore, if the designer uses the corresponding direct mechanism in which the players report their payoff types, then for any type t_i of any player i in any type space $\mathcal{T}$, reporting any $\theta_i \sim_i^f \hat{\theta}_i(t_i)$ is strictly better than reporting any $\theta_i' \not\sim_i^f \hat{\theta}_i(t_i)$, regardless of the strategies of the other players. Simply put, reporting one's true or equivalent payoff type strictly dominates reporting any other payoff type. Hence, $B_i^\infty(t_i) = \{\theta_i : \theta_i \sim_i^f \hat{\theta}_i(t_i)\}$. Thus, the direct mechanism RoRat-implements the SCF f.

Since truthful reporting one's payoff type forms an interim equilibrium on all types spaces, the direct mechanism also RoEq-implements the SCF f. Thus, semi-strict EPIC is also necessary (because it is necessary for RoRat-implementation, which is weaker than RoEq-implementation) and sufficient for RoEq-implementation in private values environments.

We thus conclude that, in private values environments, RoRat-implementation is equivalent to RoEq-implementation, and characterized by semi-strict EPIC.

7 Conclusion

We showed that RoRat-implementation is equivalent to wRat-implementation. We utilized this equivalence to prove that weak RM is necessary and almost sufficient for RoRat-implementation. We exploited the gap between weak RM and strict RM for non-responsive SCFs to establish that RoRat-implementation is strictly weaker than RoEq-implementation. We argued that the gap between RoEq-implementation and RoRat-implementation is explained by the more stringent nonemptiness requirement under the former notion, which has a bite only when the designer is allowed to use countably infinite mechanisms. An open question is whether there is any gap between the two robust implementation notions for the case of responsive SCFs even though the respective necessary conditions (strict RM and weak RM) are equivalent in this case.

8 Appendix

In the Appendix, we provide the proofs omitted from the main body of the paper.

Proof of Theorem 4.3.

It is also easy to prove this statement directly.

Suppose the mechanism $\Gamma = ((M_i)_{i\in I}, g)$ RoRat-implements f. It follows from Theorem 3.1 that Γ wRat-implements f. We now argue that f must satisfy weak RM.

Pick any $i \in I$ and $\theta \in \Theta$. Consider the belief $z_i^1 \in \Delta(\Theta_{-i})$ that puts probability one on θ_{-i}. By wRat-implementability, there exists a belief $\psi_i^\theta \in \Delta(\Theta_{-i} \times M_{-i})$ such that

(a) $\arg\max_{\tilde{m}_i \in M_i} \sum_{\tilde{\theta}_{-i}, \tilde{m}_{-i}} \psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) u_i\big(g(\tilde{m}_i, \tilde{m}_{-i}), (\theta_i, \tilde{\theta}_{-i})\big) \neq \emptyset$.

(b) $\psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) > 0 \Rightarrow \tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\tilde{\theta}_{-i})$.

(c) $\text{marg}_{\Theta_{-i}} \psi_i^\theta = z_i^1$.

If $\tilde{\theta}_{-i} \neq \theta_{-i}$, then $\psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) = 0$ because $\text{marg}_{\Theta_{-i}} \psi_i^\theta = z_i^1$ and z_i^1 assigns probability one on θ_{-i}. Therefore, for all $\tilde{m}_i \in M_i$,

$$\begin{aligned}
&\sum_{\tilde{\theta}_{-i}, \tilde{m}_{-i}} \psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) u_i\big(g(\tilde{m}_i, \tilde{m}_{-i}), (\theta_i, \tilde{\theta}_{-i})\big) \\
&= \sum_{\tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})} \text{marg}_{M_{-i}} \psi_i^\theta(\tilde{m}_{-i}) u_i\big(g(\tilde{m}_i, \tilde{m}_{-i}), \theta\big) \\
&= u_i\left(\sum_{\tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})} \text{marg}_{M_{-i}} \psi_i^\theta(\tilde{m}_{-i}) g(\tilde{m}_i, \tilde{m}_{-i}), \theta \right). \qquad (1)
\end{aligned}$$

Define the set of lotteries

$$L_i(\theta) = \left\{ \sum_{\tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})} \text{marg}_{M_{-i}} \psi_i^\theta(\tilde{m}_{-i}) g(\tilde{m}_i, \tilde{m}_{-i}) : \tilde{m}_i \in M_i \right\}.$$

Pick any $m_i \in \arg\max_{\tilde{m}_i \in M_i} \sum_{\tilde{\theta}_{-i}, \tilde{m}_{-i}} \psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) u_i\big(g(\tilde{m}_i, \tilde{m}_{-i}), (\theta_i, \tilde{\theta}_{-i})\big)$. Then $m_i \in \mathcal{S}_i^\infty(\theta_i)$ because $\psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) > 0$ implies $\tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\tilde{\theta}_{-i})$. Therefore, by wRat-implementability,

$$\sum_{\tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})} \text{marg}_{M_{-i}} \psi_i^\theta(\tilde{m}_{-i}) g(m_i, \tilde{m}_{-i}) = f(\theta).$$

Moreover, for all $\tilde{m}_i \in M_i$, we have

$$\begin{aligned}
u_i\left(\sum_{\tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})} \text{marg}_{M_{-i}} \psi_i^\theta(\tilde{m}_{-i}) g(m_i, \tilde{m}_{-i}), \theta \right) &= \sum_{\tilde{\theta}_{-i}, \tilde{m}_{-i}} \psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) u_i\big(g(m_i, \tilde{m}_{-i}), (\theta_i, \tilde{\theta}_{-i})\big) \\
&\geq \sum_{\tilde{\theta}_{-i}, \tilde{m}_{-i}} \psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) u_i\big(g(\tilde{m}_i, \tilde{m}_{-i}), (\theta_i, \tilde{\theta}_{-i})\big)
\end{aligned}$$

$$= u_i\left(\sum_{\tilde{m}_{-i}\in\mathcal{S}_{-i}^{\infty}(\theta_{-i})} \text{marg}_{M_{-i}}\psi_i^{\theta}(\tilde{m}_{-i})g(\tilde{m}_i,\tilde{m}_{-i}),\theta\right),$$

where the first and last equality follows from (1). Hence, $u_i\big(f(\theta),\theta\big) \geq u_i\big(\ell,\theta\big)$ for all $\ell \in L_i(\theta)$.

We next claim that for any $\ell \in L_i(\theta)$, $\ell \neq f(\theta)$ implies $u_i\big(f(\theta),\theta\big) > u_i\big(\ell,\theta\big)$. Suppose not. Then there is some $\ell \in L_i(\theta)$ such that $\ell \neq f(\theta)$ but $u_i\big(\ell,\theta\big) \geq u_i\big(f(\theta),\theta\big)$. By construction of $L_i(\theta)$, there exists a message $\tilde{m}_i$ such that $\sum_{\tilde{m}_{-i}\in\mathcal{S}_{-i}^{\infty}(\theta_{-i})} \text{marg}_{M_{-i}}\psi_i^{\theta}(\tilde{m}_{-i})g(\tilde{m}_i,\tilde{m}_{-i}) = \ell$. Then, as per the above arguments, $u_i(\ell,\theta) \geq u_i(f(\theta),\theta)$ is equivalent to

$$\sum_{\tilde{\theta}_{-i},\tilde{m}_{-i}} \psi_i^{\theta}(\tilde{\theta}_{-i},\tilde{m}_{-i})u_i\big(g(\tilde{m}_i,\tilde{m}_{-i}),(\theta_i,\tilde{\theta}_{-i})\big) \geq \sum_{\tilde{\theta}_{-i},\tilde{m}_{-i}} \psi_i^{\theta}(\tilde{\theta}_{-i},\tilde{m}_{-i})u_i\big(g(m_i,\tilde{m}_{-i}),(\theta_i,\tilde{\theta}_{-i})\big),$$

for some $m_i \in \arg\max_{\tilde{m}_i'\in M_i}\sum_{\tilde{\theta}_{-i},\tilde{m}_{-i}} \psi_i^{\theta}(\tilde{\theta}_{-i},\tilde{m}_{-i})u_i\big(g(\tilde{m}_i',\tilde{m}_{-i}),(\theta_i,\tilde{\theta}_{-i})\big)$. Therefore, $\tilde{m}_i$ is also a best response to the belief ψ_i^{θ} when i's payoff type is θ_i. Hence, $\tilde{m}_i \in \mathcal{S}_i^{\infty}(\theta_i)$. But $g(\tilde{m}_i,\tilde{m}_{-i}) \neq f(\theta)$ for at least one $\tilde{m}_{-i} \in \mathcal{S}_{-i}^{\infty}(\theta_{-i})$, which contradicts wRat-implementation of f.

We are now ready to prove the theorem. Consider any deception β. Define the message correspondence profile with payoff-type domain $\mathcal{S} = (\mathcal{S}_1,\ldots,\mathcal{S}_n)$ such that

$$\mathcal{S}_i(\theta_i) = \bigcup_{\theta_i'\in\beta_i(\theta_i)} \mathcal{S}_i^{\infty}(\theta_i').$$

Suppose β is not weakly refutable. Then, by definition of weak refutablility, for all $i \in I$, $\theta_i \in \Theta_i$, and $\theta_i' \in \beta_i(\theta_i)$ satisfying $\theta_i' \not\sim_i^f \theta_i$, there exist $\tilde{\theta}_i$ and $\psi_i \in \Delta(\Theta_{-i}\times\Theta_{-i})$, which satisfies $\psi_i(\theta_{-i},\theta_{-i}') > 0 \Rightarrow \theta_{-i}' \in \beta_{-i}(\theta_{-i})$, such that for all $y \in Y_i[\tilde{\theta}_i]$, we have

$$\sum_{\theta_{-i},\theta_{-i}'} \psi_i(\theta_{-i},\theta_{-i}')u_i\big(f(\theta_i',\theta_{-i}'),(\theta_i,\theta_{-i})\big) \geq \sum_{\theta_{-i},\theta_{-i}'} \psi_i(\theta_{-i},\theta_{-i}')u_i\big(y(\theta_{-i}'),(\theta_i,\theta_{-i})\big). \qquad (2)$$

We first show that for any $i \in I$, $\theta_i \in \Theta_i$, and $\theta_i' \in \beta_i(\theta_i)$ satisfying $\theta_i' \sim_i^f \theta_i$, there exist $\tilde{\theta}_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i}\times\Theta_{-i})$ satisfying $\psi_i(\theta_{-i},\theta_{-i}) > 0 \Rightarrow \theta_{-i}' \in \beta_{-i}(\theta_{-i})$ such that (2) holds for all $y \in Y_i[\tilde{\theta}_i]$.

Pick any i, θ_i, and $\theta_i' \in \beta_i(\theta_i)$ satisfying $\theta_i' \sim_i^f \theta_i$. We set $\tilde{\theta}_i = \theta_i$ and the belief $\psi_i \in \Delta(\Theta_{-i}\times\Theta_{-i})$ such that $\psi_i(\hat{\theta}_{-i},\hat{\theta}_{-i}) = 1$ for some $\hat{\theta}_{-i} \in \Theta_{-i}$. As $\hat{\theta}_{-i} \in \beta_{-i}(\hat{\theta}_{-i})$, the belief ψ_i satisfies $\psi_i(\theta_{-i},\theta_{-i}') > 0 \Rightarrow \theta_{-i}' \in \beta_{-i}(\theta_{-i})$. Since $\theta_i \sim_i^f \theta_i'$, we have $f(\theta_i',\hat{\theta}_{-i}) = f(\theta_i,\hat{\theta}_{-i})$.

Moreover, $Y_i[\tilde{\theta}_i] = Y_i[\theta_i]$ because $\tilde{\theta}_i = \theta_i$. Therefore, for all $y \in Y_i[\tilde{\theta}_i]$, we have

$$\begin{aligned}\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big) &= u_i\big(f(\theta_i,\hat{\theta}_{-i}),(\theta_i,\hat{\theta}_{-i})\big) \\ &\geq u_i\big(y(\hat{\theta}_{-i}),(\theta_i,\hat{\theta}_{-i})\big) \\ &= \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y(\theta'_{-i}),(\theta_i,\theta_{-i})\big).\end{aligned}$$

Thus, if we combine the above result with the hypothesis that β is not weakly refutable, then we can hypothesize that for all $i \in I$, $\theta_i \in \Theta_i$, and $\theta'_i \in \beta_i(\theta_i)$, there exist $\tilde{\theta}_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ satisfying $\psi_i(\theta_{-i},\theta'_{-i}) > 0 \Rightarrow \theta'_{-i} \in \beta_{-i}(\theta_{-i})$ such that (2) holds for all $y \in Y_i[\tilde{\theta}_i]$.

We next show that $b^{\Theta}(\mathcal{S}) \geq \mathcal{S}$. Pick any $i \in I$, $\theta_i \in \Theta_i$, and $m'_i \in \mathcal{S}_i(\theta_i)$. We now construct a belief $\psi_i^{\Gamma} \in \Delta(\Theta_{-i} \times M_{-i})$ satisfying $\psi_i^{\Gamma}(\theta_{-i},m_{-i}) > 0$ implies $m_{-i} \in \mathcal{S}_{-i}(\theta_{-i})$ such that m'_i is a best response for agent i of payoff type θ_i against ψ_i^{Γ}.

By definition of $\mathcal{S}$, we have $m'_i \in \mathcal{S}_i^{\infty}(\theta'_i)$ for some $\theta'_i \in \beta_i(\theta_i)$. Then, by our hypothesis, there exist $\tilde{\theta}_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ satisfying $\psi_i(\theta_{-i},\theta'_{-i}) > 0 \Rightarrow \theta'_{-i} \in \beta_{-i}(\theta_{-i})$ such that (2) holds for all $y \in Y_i[\tilde{\theta}_i]$. Define the belief $\psi_i^{\Gamma} \in \Delta(\Theta_{-i} \times M_{-i})$ as follows: for any (θ_{-i},m_{-i}),

$$\psi_i^{\Gamma}(\theta_{-i},m_{-i}) = \sum_{\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) \times \mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}).$$

By construction, $\psi_i^{\Gamma}(\theta_{-i},m_{-i}) > 0$ implies that there exists $\theta'_{-i} \in \Theta_{-i}$ such that $\psi_i(\theta_{-i},\theta'_{-i}) > 0$ and $\mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}) > 0$. But $\psi_i(\theta_{-i},\theta'_{-i}) > 0$ implies $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$. Moreover, $\mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}) > 0$ implies $m_{-i} \in \mathcal{S}_{-i}^{\infty}(\theta'_{-i})$ – recall the definition of $\psi_i^{(\tilde{\theta}_i,\theta'_{-i})}$ from the beginning of this proof. Since $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$ and $m_{-i} \in \mathcal{S}_{-i}^{\infty}(\theta'_{-i})$, it follows from the definition of $\mathcal{S}$ that $m_{-i} \in \mathcal{S}_{-i}(\theta_{-i})$.

For any $m_i \in M_i$, define $y^{m_i} : \Theta_{-i} \to \Delta(A)$ as follows: for all $\theta_{-i} \in \Theta_{-i}$,

$$y^{m_i}(\theta_{-i}) = \sum_{m_{-i}} \mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta_{-i})}(m_{-i}) g(m_i,m_{-i}).$$

By construction, $y^{m_i}(\theta_{-i}) \in L_i(\tilde{\theta}_i,\theta_{-i})$. Therefore, if $f(\tilde{\theta}_i,\theta_{-i}) \neq y^{m_i}(\theta_{-i})$, then, as argued earlier in the proof, we must have

$$u_i\big(f(\tilde{\theta}_i,\theta_{-i}),(\tilde{\theta}_i,\theta_{-i})\big) > u_i\big(y^{m_i}(\theta_{-i}),(\tilde{\theta}_i,\theta_{-i})\big).$$

So $y^{m_i} \in Y_i[\tilde{\theta}_i]$. By our hypothesis, (2) holds for all $y \in Y_i[\tilde{\theta}_i]$. Hence, for any $m_i \in M_i$,

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big) \geq \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y^{m_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big). \qquad (3)$$

We are ready to show that m'_i is a best response for agent i of payoff type θ_i against ψ_i^{Γ}.

$$\begin{aligned}
&\sum_{\theta_{-i},m_{-i}} \psi_i^{\Gamma}(\theta_{-i},m_{-i}) u_i\big(g(m'_i,m_{-i}),(\theta_i,\theta_{-i})\big) \\
=\ &\sum_{\theta_{-i},m_{-i}} \left(\sum_{\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) \times \mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}) u_i\big(g(m'_i,m_{-i}),(\theta_i,\theta_{-i})\big) \right) \\
&\quad \text{(by definition of } \psi_i^{\Gamma}) \\
=\ &\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big) \\
&\quad \left(\begin{array}{l} \text{by weak rationalizable implementability of } f \text{ because } m'_i \in \mathcal{S}_i^{\infty}(\theta'_i) \\ \text{and } \mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}) > 0 \text{ implies } m_{-i} \in \mathcal{S}_{-i}^{\infty}(\theta'_{-i}) \end{array} \right) \\
\geq\ &\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y^{m_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big) \\
&\quad (\because \text{inequality (3) holds for any } m_i \in M_i) \\
=\ &\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) \left(\sum_{m_{-i}} \mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}) u_i\big(g(m_i,m_{-i}),(\theta_i,\theta_{-i})\big) \right) \\
&\quad \text{(by definition of } y^{m_i}) \\
=\ &\sum_{\theta_{-i},m_{-i}} \psi_i^{\Gamma}(\theta_{-i},m_{-i}) u_i\big(g(m_i,m_{-i}),(\theta_i,\theta_{-i})\big) \\
&\quad \text{(by definition of } \psi_i^{\Gamma}).
\end{aligned}$$

Since m'_i is a best response for agent i of payoff type θ_i against ψ_i^{Γ} and $\psi_i^{\Gamma}(\theta_{-i},m_{-i}) > 0$ implies $m_{-i} \in \mathcal{S}_{-i}(\theta_{-i})$, it follows by definition that $m'_i \in b_i^{\Theta}(\mathcal{S})[\theta_i]$.

As $b^{\Theta}(\mathcal{S}) \geq \mathcal{S}$, we have $\mathcal{S} \leq \mathcal{S}^{\infty}$. For any $\theta \in \Theta$ and $\theta' \in \beta(\theta)$, we obtain $\mathcal{S}^{\infty}(\theta') \neq \emptyset$ since the mechanism Γ wRat-implements f. So pick any $m' \in \mathcal{S}^{\infty}(\theta') \subseteq \mathcal{S}(\theta) \subseteq \mathcal{S}^{\infty}(\theta)$. Then $g(m') = f(\theta')$ and $g(m') = f(\theta)$ because, once again, the mechanism Γ wRat-implements f. Thus, $f(\theta') = f(\theta)$. So β is acceptable. This completes the proof. □

Proof of Lemma 4.6.

Suppose the SCF f is responsive. If f satisfies strict RM, then it clearly satisfies weak RM.

Now, suppose f satisfies weak RM. Fix an unacceptable deception β. Then β is weakly

refutable. Thus, there exist $i \in I$, $\theta_i \in \Theta_i$, and $\theta_i^{'} \in \beta_i(\theta_i)$ satisfying $\theta_i^{'} \not\sim_i^f \theta_i$ such that for all $\tilde{\theta}_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ satisfying $\psi_i(\theta_{-i}, \theta_{-i}^{'}) > 0 \Rightarrow \theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$, there exists $y \in Y_i[\tilde{\theta}_i]$ such that

$$\sum_{\theta_{-i}, \theta_{-i}^{'}} \psi_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(y(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta_{-i}^{'}} \psi_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(f(\theta_i^{'}, \theta_{-i}^{'}), (\theta_i, \theta_{-i})\big).$$

Pick any belief $\hat{\psi}_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ satisfying $\hat{\psi}_i(\theta_{-i}, \theta_{-i}^{'}) > 0 \Rightarrow \theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$. Then, for $\theta_i^{'}$, there exists $y^{'} \in Y_i[\theta_i^{'}]$ such that

$$\sum_{\theta_{-i}, \theta_{-i}^{'}} \hat{\psi}_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(y^{'}(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta_{-i}^{'}} \hat{\psi}_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(f(\theta_i^{'}, \theta_{-i}^{'}), (\theta_i, \theta_{-i})\big).$$

Pick any $\epsilon \in (0, 1)$ and define $y^{\epsilon} : \Theta_{-i} \to \Delta(A)$ such that, for any $\theta_{-i} \in \Theta_{-i}$,

$$y^{\epsilon}(\theta_{-i}) = \epsilon y^{'}(\theta_{-i}) + (1 - \epsilon) f(\theta_i^{'}, \theta_{-i}).$$

As f is responsive, if $\tilde{\theta}_i \neq \theta_i^{'}$, then $\tilde{\theta}_i \not\sim_i^f \theta_i^{'}$. Moreover, since f satisfies weak RM, it satisfies semi-strict EPIC (Lemma 5.4). Hence, if $\tilde{\theta}_i \neq \theta_i^{'}$, then $u_i\big(f(\tilde{\theta}_i, \theta_{-i}), (\tilde{\theta}_i, \theta_{-i})\big) > u_i\big(f(\theta_i^{'}, \theta_{-i}), (\tilde{\theta}_i, \theta_{-i})\big)$ for all θ_{-i}. Since Θ is finite, we can find a sufficiently small ϵ such that $u_i\big(f(\tilde{\theta}_i, \theta_{-i}), (\tilde{\theta}_i, \theta_{-i})\big) > u_i\big(y^{\epsilon}(\theta_{-i}), (\tilde{\theta}_i, \theta_{-i})\big)$ for all θ_{-i} and $\tilde{\theta}_i \neq \theta_i^{'}$. Thus, $y^{\epsilon} \in Y_i[\tilde{\theta}_i]$ for all $\tilde{\theta}_i \neq \theta_i^{'}$. Moreover, $y^{\epsilon} \in Y_i[\theta_i^{'}]$ since both $y^{'}$ and $f(\theta_i^{'}, \cdot)$ are in $Y_i[\theta_i^{'}]$. We thus conclude that $y^{\epsilon} \in \bigcap_{\tilde{\theta}_i \in \Theta_i} Y_i[\tilde{\theta}_i]$.

Since ϵ is positive, by construction of y^{ϵ}, we have

$$\sum_{\theta_{-i}, \theta_{-i}^{'}} \hat{\psi}_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(y^{\epsilon}(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta_{-i}^{'}} \hat{\psi}_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(f(\theta_i^{'}, \theta_{-i}^{'}), (\theta_i, \theta_{-i})\big).$$

Therefore, β is strictly refutable. Hence, f satisfies strict RM. □

Proof of Lemma 5.2.

We prove (a) first. Suppose the SCF f satisfies conditional NTI. Pick any $i \in I$, $\theta_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$.

Firstly, it follows from the definition of conditional NTI that for all $\theta_{-i}^{'} \in \Theta_{-i}$, there exists $\ell^{\theta_{-i}^{'}} \in \Delta(A)$ such that

$$u_i\big(f(\theta_i, \theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) > u_i\big(\ell^{\theta_{-i}^{'}}, (\theta_i, \theta_{-i}^{'})\big). \tag{4}$$

To see this, consider the degenerate belief $\tilde{\psi}_i$ such that $\tilde{\psi}_i(\theta_{-i}^{'}, \theta_{-i}^{'}) = 1$. Then there must

exist $\tilde{y}, \tilde{y}^{'} \in Y_i^w[\theta_i]$ such that

$$\begin{aligned} u_i\big(f(\theta_i, \theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) &\geq u_i\big(\tilde{y}(\theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) \\ &= \sum_{\theta_{-i}, \theta_{-i}^{''}} \tilde{\psi}_i(\theta_{-i}, \theta_{-i}^{''}) u_i\big(\tilde{y}(\theta_{-i}^{''}), (\theta_i, \theta_{-i})\big) \\ &> \sum_{\theta_{-i}, \theta_{-i}^{''}} \tilde{\psi}_i(\theta_{-i}, \theta_{-i}^{''}) u_i\big(\tilde{y}^{'}(\theta_{-i}^{''}), (\theta_i, \theta_{-i})\big) \\ &= u_i\big(\tilde{y}^{'}(\theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big), \end{aligned}$$

where the first weak inequality follows from the fact that $\tilde{y} \in Y_i^w[\theta_i]$ and the strict inequality follows from conditional NTI. Then $\ell^{\theta_{-i}^{'}} = \tilde{y}^{'}(\theta_{-i}^{'})$ satisfies (4).

Secondly, since f satisfies conditional NTI, there exist $y, y^{'} \in Y_i^w[\theta_i]$ such that

$$\sum_{\theta_{-i}, \theta_{-i}^{'}} \psi_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(y(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta_{-i}^{'}} \psi_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(y^{'}(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big).$$

Pick any $\epsilon \in (0, 1)$ and define $y^{\epsilon} : \Theta_{-i} \to \Delta(A)$ such that $y^{\epsilon}(\theta_{-i}^{'}) = (1-\epsilon) y(\theta_{-i}^{'}) + \epsilon \ell^{\theta_{-i}^{'}}$ for all $\theta_{-i}^{'}$. We similarly define $y^{'\epsilon}$. By construction, y^{ϵ} and $y^{'\epsilon}$ are such that for all $\theta_{-i}^{'} \in \Theta_{-i}$,

$$u_i\big(f(\theta_i, \theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) > u_i\big(y^{\epsilon}(\theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) \text{ and } u_i\big(f(\theta_i, \theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) > u_i\big(y^{'\epsilon}(\theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big).$$

For ϵ sufficiently close to 1, we have

$$\sum_{\theta_{-i}, \theta_{-i}^{'}} \psi_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(y^{\epsilon}(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta_{-i}^{'}} \psi_i(\theta_{-i}, \theta_{-i}^{'}) u_i\big(y^{'\epsilon}(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big).$$

We fix any such sufficiently large ϵ.

Thirdly, since $\Delta^*(A)$ is a dense subset of $\Delta(A)$, for each $\theta_{-i}^{'}$, there exists a sequence of lotteries $\{\ell^z(\theta_{-i}^{'})\}_{z=1}^{\infty} \in \Delta^*(A)$ converging to $y^{\epsilon}(\theta_{-i}^{'})$. For each $z \geq 1$, define $y^z : \Theta_{-i} \to \Delta^*(A)$ such that $y^z(\theta_{-i}^{'}) = \ell^z(\theta_{-i}^{'})$ for all $\theta_{-i}^{'}$. Similarly, we can define $y^{'z} : \Theta_{-i} \to \Delta^*(A)$ such that $y^{'z}(\theta_{-i}^{'})$ converges to $y^{'\epsilon}(\theta_{-i}^{'})$ for all $\theta_{-i}^{'}$. As Θ_{-i} is finite, there exists a sufficiently large z such that

$$u_i\big(f(\theta_i, \theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) > u_i\big(y^{z}(\theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) \text{ and } u_i\big(f(\theta_i, \theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big) > u_i\big(y^{'z}(\theta_{-i}^{'}), (\theta_i, \theta_{-i}^{'})\big),$$

for all θ'_{-i}, and

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y^z(\theta'_{-i}),(\theta_i,\theta_{-i})\big) > \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y'^z(\theta'_{-i}),(\theta_i,\theta_{-i})\big). \qquad (5)$$

The first set of inequalities imply that $y^z, y'^z \in Y_i^*[\theta_i]$.

Lastly, since $y_i^{\theta_i}$ assigns a positive weight to all $y \in Y_i^*[\theta_i]$, if

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y_i^{\theta_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big) \geq \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y(\theta'_{-i}),(\theta_i,\theta_{-i})\big), \forall y \in Y_i^*[\theta_i],$$

then it must be that

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y^z(\theta'_{-i}),(\theta_i,\theta_{-i})\big) = \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y'^z(\theta'_{-i}),(\theta_i,\theta_{-i})\big),$$

which contradicts (5).

We prove (b) next. Suppose the SCF f satisfies conditional NTI. Pick any $i \in I$, $\theta_i \in \Theta_i$ and $z_i^1 \in \Delta(\Theta_{-i})$. As $\bar{\alpha}$ assigns a positive weight to all $a \in A$, if

$$\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(\bar{\alpha},(\theta_i,\theta_{-i})\big) \geq \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(a,(\theta_i,\theta_{-i})\big), \forall a \in A,$$

then it must be that

$$\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(a,(\theta_i,\theta_{-i})\big) = \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(a',(\theta_i,\theta_{-i})\big),$$

for all $a, a' \in A$. Now consider the belief $\tilde{\psi}_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ such that $\tilde{\psi}_i(\theta_{-i},\theta_{-i}) = z_i^1(\theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$. Then, by conditional NTI, there must exist $\tilde{y}, \tilde{y}' \in Y_i^w[\theta_i]$ such that

$$\sum_{\theta_{-i},\theta'_{-i}} \tilde{\psi}_i(\theta_{-i},\theta'_{-i}) u_i\big(\tilde{y}(\theta'_{-i}),(\theta_i,\theta_{-i})\big) > \sum_{\theta_{-i},\theta'_{-i}} \tilde{\psi}_i(\theta_{-i},\theta'_{-i}) u_i\big(\tilde{y}'(\theta'_{-i}),(\theta_i,\theta_{-i})\big).$$

But the left-hand side of the above inequality equals $\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(\tilde{y}(\theta_{-i}),(\theta_i,\theta_{-i})\big)$ while the right-hand side equals $\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(\tilde{y}'(\theta_{-i}),(\theta_i,\theta_{-i})\big)$, which contradicts the fact that type θ_i is indifferent over all alternatives when she holds the belief z_i^1. □

Proof of Lemma 5.4.

Suppose the SCF f satisfies weak RM. Pick any $i \in I$, $\theta_i, \theta'_i \in \Theta_i$. If $\theta_i \sim_i^f \theta'_i$, then trivially

$u_i\big(f(\theta_i,\theta_{-i}),(\theta_i,\theta_{-i})\big) \geq u_i\big(f(\theta_i^{'},\theta_{-i}),(\theta_i,\theta_{-i})\big)$ for all $\theta_{-i} \in \Theta_{-i}$. So suppose $\theta_i \not\sim_i^f \theta_i^{'}$. Consider the deception β such that $\beta_j(\theta_j) = \{\theta_j\}$ for all θ_j and $j \neq i$ but

$$\beta_i(\tilde{\theta}_i) = \begin{cases} \{\theta_i, \theta_i^{'}\}, & \text{if } \tilde{\theta}_i = \theta_i \\ \{\tilde{\theta}_i\}, & \text{otherwise.} \end{cases}$$

Since $\theta_i \not\sim_i^f \theta_i^{'}$, the deception β is unacceptable. Hence, it must be weakly refutable. That is, there exist $j \in I$, $\hat{\theta}_j \in \Theta_j$, and $\hat{\theta}_j^{'} \in \beta_j(\hat{\theta}_j)$ satisfying $\hat{\theta}_j^{'} \not\sim_j^f \hat{\theta}_j$ such that for any $\tilde{\theta}_j \in \Theta_j$ and $\psi_j \in \Delta(\Theta_{-j} \times \Theta_{-j})$ satisfying $\psi_j(\theta_{-j}, \theta_{-j}^{'}) > 0 \Rightarrow \theta_{-j}^{'} \in \beta_{-j}(\theta_{-j})$, there exists $y \in Y_j[\tilde{\theta}_j]$ such that

$$\sum_{\theta_{-j},\theta_{-j}^{'}} \psi_j(\theta_{-j},\theta_{-j}^{'}) u_j\big(y(\theta_{-j}^{'}),(\hat{\theta}_j,\theta_{-j})\big) > \sum_{\theta_{-j},\theta_{-j}^{'}} \psi_j(\theta_{-j},\theta_{-j}^{'}) u_j\big(f(\hat{\theta}_j^{'},\theta_{-j}^{'}),(\hat{\theta}_j,\theta_{-j})\big).$$

Since $\hat{\theta}_j^{'} \not\sim_j^f \hat{\theta}_j$ and $\hat{\theta}_j^{'} \in \beta_j(\hat{\theta}_j)$, it must be that $j = i$, $\hat{\theta}_j = \theta_i$ and $\hat{\theta}_j^{'} = \theta_i^{'}$.

Now pick any $\theta_{-i} \in \Theta_{-i}$. Consider $\tilde{\theta}_i = \theta_i$ and the degenerate belief ψ_i such that $\psi_i(\theta_{-i},\theta_{-i}) = 1$. Note that $\theta_{-i} \in \beta_{-i}(\theta_{-i})$. Hence, we must have some $y \in Y_i[\tilde{\theta}_i] = Y_i[\theta_i]$ such that $u_i\big(y(\theta_{-i}),(\theta_i,\theta_{-i})\big) > u_i\big(f(\theta_i^{'},\theta_{-i}),(\theta_i,\theta_{-i})\big)$. But $y \in Y_i[\theta_i]$ implies that $u_i\big(f(\theta_i,\theta_{-i}),(\theta_i,\theta_{-i})\big) \geq u_i\big(y(\theta_{-i}),(\theta_i,\theta_{-i})\big)$. We thus conclude that $u_i\big(f(\theta_i,\theta_{-i}),(\theta_i,\theta_{-i})\big) > u_i\big(f(\theta_i^{'},\theta_{-i}),(\theta_i,\theta_{-i})\big)$. □

Proof of Theorem 5.5.

We use the mechanism Γ constructed above and prove that Γ wRat-implements f, which implies that Γ RoRat-implements f because of Theorem 3.1. The proof of the theorem consists of Steps 1 through 4.

Step 1: $m_i \in \mathcal{S}_i^{\infty}(\theta_i) \Rightarrow m_i^2 = 1$.

Proof. Suppose by way of contradiction that $m_i \in \mathcal{S}_i^{\infty}(\theta_i)$ but $m_i^2 > 1$. Then, m_i is a best response of individual i of payoff type θ_i against some conjecture $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$.

For each $\theta_i^{'} \neq \theta_i^*$ and $\theta_{-i}^{'} \in \Theta_{-i}$, we define

$$M_{-i}^2(\theta_i^{'},\theta_{-i}^{'}) \;=\; \Big\{m_{-i} : m_j^2 = 1 \text{ and } m_j^1[i] = \theta_i^{'}, \forall j \neq i, \text{and } (m_j^1[j])_{j\neq i} = \theta_{-i}^{'}\Big\}.$$

For θ_i^* and each $\theta_{-i}^{'} \in \Theta_{-i}$, we define

$$M_{-i}^2(\theta_i^*,\theta_{-i}^{'}) \;=\; \left\{ m_{-i} : \begin{array}{l} (m_j^1[j])_{j\neq i} = \theta_{-i}^{'} \text{ and} \\ \text{either } m_j^2 = 1 \text{ and } m_j^1[i] = \theta_i^*, \forall j \neq i, \\ \text{or } m_j^2 = 1, \forall j \neq i, \text{ but } m_{j^{'}}^1[i] \neq m_k^1[i] \text{ for some } j^{'}, k \neq i \end{array} \right\}.$$

Also define

$$M_{-i}^3 \;=\; \left\{m_{-i} : \text{there exist one or more } j \neq i \text{ such that } m_j^2 > 1\right\}.$$

Note that $\left((M_{-i}^2(\tilde{\theta}_i, \theta'_{-i}))_{\tilde{\theta}_i \in \Theta_i, \theta'_{-i} \in \Theta_{-i}}, M_{-i}^3\right)$ defines a partition of M_{-i}. As $m_i^2 > 1$, if $m_{-i} \in M_{-i}^2(\tilde{\theta}_i, \theta'_{-i})$, then Rule 2 is used under the profile (m_i, m_{-i}) whereas if $m_{-i} \in M_{-i}^3$, then Rule 3 is used under the profile (m_i, m_{-i}).

For each $\tilde{\theta}_i \in \Theta_i$, define

$$\Psi_i^{2,\tilde{\theta}_i} = \sum_{\theta_{-i}, \theta''_{-i}} \sum_{m_{-i} \in M_{-i}^2(\tilde{\theta}_i, \theta''_{-i})} \psi_i(\theta_{-i}, m_{-i}).$$

Thus, $\Psi_i^{2,\tilde{\theta}_i}$ is the probability of the event that all other individuals report a message profile in $\bigcup_{\theta''_{-i}} M_{-i}^2(\tilde{\theta}_i, \theta''_{-i})$.

Also, define

$$\Psi_i^3 = \sum_{\theta_{-i}, m_{-i} \in M_{-i}^3} \psi_i(\theta_{-i}, m_{-i}).$$

Thus, Ψ_i^3 is the probability of the event that all other individuals report a message profile in M_{-i}^3.

If $\tilde{\theta}_i$ is such that $\Psi_i^{2,\tilde{\theta}_i} > 0$, then define $\psi_i^{2,\tilde{\theta}_i} \in \Delta(\Theta_{-i} \times \Theta_{-i})$ such that for all $\theta_{-i}, \theta'_{-i} \in \Theta_{-i}$,

$$\psi_i^{2,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) = \sum_{m_{-i} \in M_{-i}^2(\tilde{\theta}_i, \theta'_{-i})} \frac{\psi_i(\theta_{-i}, m_{-i})}{\Psi_i^{2,\tilde{\theta}_i}}.$$

Thus, $\psi_i^{2,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i})$ is the conditional probability of the event that the payoff-type profile of all other individuals is θ_{-i} and they report a message profile in $M_{-i}^2(\tilde{\theta}_i, \theta'_{-i})$ given the event that all other individuals report a message profile in $\bigcup_{\theta''_{-i}} M_{-i}^2(\tilde{\theta}_i, \theta''_{-i})$.

If the payoff-type profile of all other individuals is θ_{-i} and they report a message profile in $M_{-i}^2(\tilde{\theta}_i, \theta'_{-i})$, then when individual i of payoff type θ_i plays m_i, she expects the outcome to be given by the lottery

$$\left(\frac{m_i^2}{1+m_i^2}\right) m_i^3[\tilde{\theta}_i](\theta'_{-i}) + \left(1 - \frac{m_i^2}{1+m_i^2}\right) y_i^{\tilde{\theta}_i}(\theta'_{-i}).$$

As a result, conditional on the event that all other individuals report a message profile in

$\bigcup_{\theta''_{-i}} M^2_{-i}(\tilde{\theta}_i, \theta''_{-i})$, the expected payoff of individual i of payoff type θ_i when she plays m_i is

$$\left(\frac{m_i^2}{1+m_i^2}\right) \sum_{\theta_{-i}, \theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) u_i\big(m_i^3[\tilde{\theta}_i](\theta'_{-i}), (\theta_i, \theta_{-i})\big)$$
$$+ \left(1 - \frac{m_i^2}{1+m_i^2}\right) \sum_{\theta_{-i}, \theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) u_i\big(y_i^{\tilde{\theta}_i}(\theta'_{-i}), (\theta_i, \theta_{-i})\big). \tag{6}$$

If $\Psi_i^3 > 0$, then define $\psi_i^3 \in \Delta(\Theta_{-i})$ such that, for any $\theta_{-i} \in \Theta_{-i}$,

$$\psi_i^3(\theta_{-i}) = \sum_{m_{-i} \in M^3_{-i}} \frac{\psi_i(\theta_{-i}, m_{-i})}{\Psi_i^3}.$$

Thus, $\psi_i^3(\theta_{-i})$ is the conditional probability of the event that the payoff-type profile of all other individuals is θ_{-i} and they report a message profile in M^3_{-i} given the event that all other individuals report a message profile in M^3_{-i}.

If the payoff-type profile of all other individuals is θ_{-i} and they report a message profile $m_{-i} \in M^3_{-i}$, then when individual i of payoff type θ_i plays m_i, she expects the outcome to be given by the lottery

$$\frac{1}{n}\left(\frac{m_i^2}{1+m_i^2}\right) m_i^4 + \frac{1}{n}\left(1 - \frac{m_i^2}{1+m_i^2}\right) \bar{\alpha} + \sum_{j \neq i} \left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right) m_j^4 + \frac{1}{n}\left(1 - \frac{m_j^2}{1+m_j^2}\right) \bar{\alpha}\right).$$

As a result, conditional on the event that all other individuals report a message profile in M^3_{-i}, the expected payoff of individual i of payoff type θ_i when she plays m_i is

$$\frac{1}{n}\left(\frac{m_i^2}{1+m_i^2}\right) \sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(m_i^4, (\theta_i, \theta_{-i})\big) + \frac{1}{n}\left(1 - \frac{m_i^2}{1+m_i^2}\right) \sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(\bar{\alpha}, (\theta_i, \theta_{-i})\big)$$
$$+ \sum_{\theta_{-i}, m_{-i} \in M^3_{-i}} \frac{\psi_i(\theta_{-i}, m_{-i})}{\Psi_i^3} \sum_{j \neq i} \left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right) u_i\big(m_j^4, (\theta_i, \theta_{-i})\big) + \frac{1}{n}\left(1 - \frac{m_j^2}{1+m_j^2}\right) u_i\big(\bar{\alpha}, (\theta_i, \theta_{-i})\big)\right). \tag{7}$$

Now let individual i of payoff type θ_i deviate to $\hat{m}_i = (m_i^1, \hat{m}_i^2, \hat{m}_i^3, \hat{m}_i^4)$ such that

- $\hat{m}_i^2 = m_i^2 + 1$.
- $\hat{m}_i^3$ is defined as follows: for each $\tilde{\theta}_i$:

▷ If $\Psi_i^{2,\tilde{\theta}_i} > 0$, then let $\hat{m}_i^3[\tilde{\theta}_i] \in Y_i^*[\tilde{\theta}_i]$ be such that

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(\hat{m}_i^3[\tilde{\theta}_i](\theta'_{-i}),(\theta_i,\theta_{-i})\big) \geq \sum_{\theta_{-i},\theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(m_i^3[\tilde{\theta}_i](\theta'_{-i}),(\theta_i,\theta_{-i})\big)$$

and

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(\hat{m}_i^3[\tilde{\theta}_i](\theta'_{-i}),(\theta_i,\theta_{-i})\big) > \sum_{\theta_{-i},\theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(y_i^{\tilde{\theta}_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big).$$

Note that such $\hat{m}_i^3[\tilde{\theta}_i]$ exists because of Lemma 5.2.

▷ If $\Psi_i^{2,\tilde{\theta}_i} = 0$, then let $\hat{m}_i^3[\tilde{\theta}_i] = m_i^3[\tilde{\theta}_i]$.

- $\hat{m}_i^4$ is defined as follows:

  ▷ If $\Psi_i^3 > 0$, then let $\hat{m}_i^4 \in A$ be such that

  $$\sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(\hat{m}_i^4,(\theta_i,\theta_{-i})\big) \geq \sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(m_i^4,(\theta_i,\theta_{-i})\big)$$

  and

  $$\sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(\hat{m}_i^4,(\theta_i,\theta_{-i})\big) > \sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(\bar{\alpha},(\theta_i,\theta_{-i})\big).$$

  Note that such $\hat{m}_i^4$ exists because of Lemma 5.2.

  ▷ If $\Psi_i^3 = 0$, then let $\hat{m}_i^4 = m_i^4$.

If $\Psi_i^{2,\tilde{\theta}_i} > 0$, then conditional on the event that all other individuals report a message profile in $\bigcup_{\theta''_{-i}} M_{-i}^2(\tilde{\theta}_i,\theta''_{-i})$, the expected payoff of individual i of payoff type θ_i when she plays $\hat{m}_i$ is

$$\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) \sum_{\theta_{-i},\theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(\hat{m}_i^3[\tilde{\theta}_i](\theta'_{-i}),(\theta_i,\theta_{-i})\big)$$
$$+ \left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) \sum_{\theta_{-i},\theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(y_i^{\tilde{\theta}_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big),$$

which is, by construction, greater than her expected payoff in (6) when she plays m_i.

If $\Psi_i^3 > 0$, then conditional on the event that all other individuals report a message profile in M_{-i}^3, the expected payoff of individual i of payoff type θ_i when she plays $\hat{m}_i$ is

$$\frac{1}{n}\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) \sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(\hat{m}_i^4,(\theta_i,\theta_{-i})\big) + \frac{1}{n}\left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) \sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(\bar{\alpha},(\theta_i,\theta_{-i})\big)$$

$$+\sum_{\theta_{-i},m_{-i}\in M^3_{-i}}\frac{\psi_i(\theta_{-i},m_{-i})}{\Psi_i^3}\sum_{j\neq i}\left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right)u_i\big(m_j^4,(\theta_i,\theta_{-i})\big)+\frac{1}{n}\left(1-\frac{m_j^2}{1+m_j^2}\right)u_i\big(\bar{\alpha},(\theta_i,\theta_{-i})\big)\right),$$

which is, by construction, greater than her expected payoff in (7) when she plays m_i.

As $\sum_{\tilde{\theta}_i}\Psi_i^{2,\tilde{\theta}_i}+\Psi_i^3=1$ (because $m_i^2>1$), it follows that $\hat{m}_i$ is a better response for individual i of payoff type θ_i against ψ_i, a contradiction. This completes the proof of Step 1. □

Step 2: For each $i\in I$ and $\theta_i\in\Theta_i$, let

$$\beta_i(\theta_i)=\{\theta_i\}\cup\{\theta_i^{'}\in\Theta_i:\exists\ m_i\in\mathcal{S}_i^{\infty}(\theta_i)\text{ such that }m_i^1[i]=\theta_i^{'}\}.$$

Then, the deception $\beta=(\beta_i)_{i\in I}$ is acceptable.

Proof. Suppose not, that is, β is unacceptable. Then, by weak RM, β must be weakly refutable. That is, there exist $i\in I$, $\theta_i\in\Theta_i$, and $\theta_i^{'}\in\beta_i(\theta_i)$ satisfying $\theta_i^{'}\not\succsim_i^f\theta_i$ such that for all $\tilde{\theta}_i\in\Theta_i$ and $\psi_i\in\Delta(\Theta_{-i}\times\Theta_{-i})$ satisfying $\psi_i(\theta_{-i},\theta_{-i}^{'})>0\Rightarrow\theta_{-i}^{'}\in\beta_{-i}(\theta_{-i})$, there exists $y\in Y_i[\tilde{\theta}_i]$ such that

$$\sum_{\theta_{-i},\theta_{-i}^{'}}\psi_i(\theta_{-i},\theta_{-i}^{'})u_i\big(y(\theta_{-i}^{'}),(\theta_i,\theta_{-i})\big)>\sum_{\theta_{-i},\theta_{-i}^{'}}\psi_i(\theta_{-i},\theta_{-i}^{'})u_i\big(f(\theta_i^{'},\theta_{-i}^{'}),(\theta_i,\theta_{-i})\big).$$

As $\theta_i^{'}\not\succsim_i^f\theta_i$ and $\theta_i^{'}\in\beta_i(\theta_i)$, we can find a message $m_i\in\mathcal{S}_i^{\infty}(\theta_i)$ such that $m_i^1[i]=\theta_i^{'}$. Then, m_i is a best response to some belief $\psi_i^{\Gamma}\in\Delta(\Theta_{-i}\times M_{-i})$ such that $\psi_i^{\Gamma}(\theta_{-i},m_{-i})>0\Rightarrow m_{-i}\in\mathcal{S}_{-i}^{\infty}(\theta_{-i})$. From Step 1, it follows that $\psi_i^{\Gamma}(\theta_{-i},m_{-i})>0$ implies $m_j^2=1$ for all $j\neq i$. We next define a partition of all those message profiles in M_{-i} such that $m_j^2=1$ for all $j\neq i$.

For each $\hat{\theta}_i\neq\theta_i^*$ and $\theta_{-i}^{'}\in\Theta_{-i}$, we define

$$M_{-i}^1(\hat{\theta}_i,\theta_{-i}^{'})\ =\ \left\{m_{-i}:m_j^2=1\text{ and }m_j^1[i]=\hat{\theta}_i,\forall j\neq i,\text{and }(m_j^1[j])_{j\neq i}=\theta_{-i}^{'}\right\}.$$

For θ_i^* and each $\theta_{-i}^{'}\in\Theta_{-i}$, we define

$$M_{-i}^1(\theta_i^*,\theta_{-i}^{'})\ =\ \left\{m_{-i}:\begin{array}{l}(m_j^1[j])_{j\neq i}=\theta_{-i}^{'}\text{ and}\\ \text{either }m_j^2=1\text{ and }m_j^1[i]=\theta_i^*,\forall j\neq i,\\ \text{or }m_j^2=1,\forall j\neq i,\text{ but }m_{j^{'}}^1[i]\neq m_k^1[i]\text{ for some }j^{'},k\neq i\end{array}\right\}.$$

For each $\tilde{\theta}_i \in \Theta_i$, we define

$$\Psi_i^{1,\tilde{\theta}_i} = \sum_{\theta_{-i},\theta''_{-i}} \sum_{m_{-i} \in M^1_{-i}(\tilde{\theta}_i,\theta''_{-i})} \psi_i^{\Gamma}(\theta_{-i}, m_{-i}).$$

Thus, $\Psi_i^{1,\tilde{\theta}_i}$ is the probability of the event that all other individuals report a message profile in $\bigcup_{\theta''_{-i}} M^1_{-i}(\tilde{\theta}_i, \theta''_{-i})$.

If $\tilde{\theta}_i$ is such that $\Psi_i^{1,\tilde{\theta}_i} > 0$, then define $\psi_i^{1,\tilde{\theta}_i} \in \Delta(\Theta_{-i} \times \Theta_{-i})$ such that for all $\theta_{-i}, \theta'_{-i} \in \Theta_{-i}$,

$$\psi_i^{1,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) = \sum_{m_{-i} \in M^1_{-i}(\tilde{\theta}_i,\theta'_{-i})} \frac{\psi_i^{\Gamma}(\theta_{-i}, m_{-i})}{\Psi_i^{1,\tilde{\theta}_i}}.$$

Thus, $\psi_i^{1,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i})$ is the conditional probability of the event that the payoff-type profile of all other individuals is θ_{-i} and they report a message profile in $M^1_{-i}(\tilde{\theta}_i, \theta'_{-i})$ given the event that all other individuals report a message profile in $\bigcup_{\theta''_{-i}} M^1_{-i}(\tilde{\theta}_i, \theta''_{-i})$.

If the payoff-type profile of all other individuals is θ_{-i} and they report a message profile in $M^1_{-i}(\tilde{\theta}_i, \theta'_{-i})$, then when individual i of payoff type θ_i plays m_i, she expects the outcome to be $f(\theta'_i, \theta'_{-i})$. As a result, conditional on the event that all other individuals report a message profile in $\bigcup_{\theta''_{-i}} M^1_{-i}(\tilde{\theta}_i, \theta''_{-i})$, the expected payoff of individual i of payoff type θ_i when she plays m_i is

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i^{1,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) u_i\big(f(\theta'_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big). \tag{8}$$

Now, $\psi_i^{1,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) > 0$ implies that $\psi_i^{\Gamma}(\theta_{-i}, m_{-i}) > 0$ for some $m_{-i} \in M^1_{-i}(\tilde{\theta}_i, \theta'_{-i})$. But $\psi_i^{\Gamma}(\theta_{-i}, m_{-i}) > 0$ also implies that $m_{-i} \in \mathcal{S}^{\infty}_{-i}(\theta_{-i})$. Hence, due to the construction of β, we have $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$. So, it follows from weak refutability of β that there exists $y[\tilde{\theta}_i] \in Y_i[\tilde{\theta}_i]$ such that

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i^{1,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) u_i\big(y[\tilde{\theta}_i](\theta'_{-i}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i},\theta'_{-i}} \psi_i^{1,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) u_i\big(f(\theta'_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big).$$

It is without loss of generality to assume that $y[\tilde{\theta}_i] \in Y_i^*[\tilde{\theta}_i]$. If not, then consider any sequence $\ell^z : \Theta_{-i} \to \Delta^*(A) \cup \{f(\tilde{\theta}_i, \theta_{-i})\}$ such that (a) if $y[\tilde{\theta}_i](\theta_{-i}) = f(\tilde{\theta}_i, \theta_{-i})$, then $\ell^z(\theta_{-i}) = f(\tilde{\theta}_i, \theta_{-i})$ for all $z \in \mathbb{N}$ and (b) if $y[\tilde{\theta}_i](\theta_{-i}) \neq f(\tilde{\theta}_i, \theta_{-i})$, then $\ell^z(\theta_{-i})$ converges to $y[\tilde{\theta}_i](\theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$ as $z \to \infty$. As Θ_{-i} is finite and $u_i(\cdot, \theta)$ is continuous over $\Delta(A)$,

we can find a sufficiently large $\hat{z}$ such that

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i^{1,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(\ell^{\hat{z}}(\theta'_{-i}),(\theta_i,\theta_{-i})\big) > \sum_{\theta_{-i},\theta'_{-i}} \psi_i^{1,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big),$$

and, because $y[\tilde{\theta}_i] \in Y_i[\tilde{\theta}_i]$, if $\ell^{\hat{z}}(\theta_{-i}) \neq f(\tilde{\theta}_i,\theta_{-i})$, then

$$u_i\big(f(\tilde{\theta}_i,\theta_{-i}),(\tilde{\theta}_i,\theta_{-i})\big) > u_i\big(\ell^{\hat{z}}(\theta_{-i}),(\tilde{\theta}_i,\theta_{-i})\big).$$

The latter condition implies that $\ell^{\hat{z}} \in Y_i^*[\tilde{\theta}_i]$.

Now, let individual i of payoff type θ_i deviate to $\hat{m}_i = (m_i^1, \hat{m}_i^2, \hat{m}_i^3, m_i^4)$ such that

- $\hat{m}_i^2 > 1$, where the specific value is chosen later.
- $\hat{m}_i^3$ is defined as follows: for each $\tilde{\theta}_i \in \Theta_i$:
  - ▷ If $\Psi_i^{1,\tilde{\theta}_i} > 0$, then let $\hat{m}_i^3[\tilde{\theta}_i] = y[\tilde{\theta}_i]$.
  - ▷ If $\Psi_i^{1,\tilde{\theta}_i} = 0$, then let $\hat{m}_i^3[\tilde{\theta}_i] = m_i^3[\tilde{\theta}_i]$.

If $\Psi_i^{1,\tilde{\theta}_i} > 0$, then conditional on the event that all other individuals report a message profile in $\bigcup_{\theta''_{-i}} M_{-i}^1(\tilde{\theta}_i,\theta''_{-i})$, the expected payoff of individual i of payoff type θ_i when she plays $\hat{m}_i$ is

$$\begin{aligned}
&\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) \sum_{\theta_{-i},\theta'_{-i}} \psi_i^{1,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(y[\tilde{\theta}_i](\theta'_{-i}),(\theta_i,\theta_{-i})\big) \\
&+ \left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) \sum_{\theta_{-i},\theta'_{-i}} \psi_i^{1,\tilde{\theta}_i}(\theta_{-i},\theta'_{-i}) u_i\big(y_i^{\tilde{\theta}_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big).
\end{aligned}$$

If $\hat{m}_i^2$ is large enough, then the above expression is greater than her expected payoff in (8) when she plays m_i. Since Θ_i is finite, we can find a sufficiently large $\hat{m}_i^2$ such that the above statement is true for all $\tilde{\theta}_i \in \Theta_i$ such that $\Psi_i^{1,\tilde{\theta}_i} > 0$. As $\sum_{\tilde{\theta}_i} \Psi_i^{1,\tilde{\theta}_i} = 1$ (because $\psi_i^\Gamma(\theta_{-i},m_{-i}) > 0 \Rightarrow m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i}) \Rightarrow m_j^2 = 1, \forall j \neq i$), it follows that $\hat{m}_i$ is a better response for individual i of payoff type θ_i against ψ_i^Γ, a contradiction. This completes the proof of Step 2. □

It follows from Steps 1 and 2 that $m \in \mathcal{S}^\infty(\theta) \Rightarrow g(m) = f(\theta)$.

Step 3: Define the message correspondence profile with payoff-type domain $\mathcal{S} = (\mathcal{S}_1,\ldots,\mathcal{S}_n)$

such that for all $i \in I$ and $\theta_i \in \Theta_i$,

$$\mathcal{S}_i(\theta_i) = \{(m_i^1, 1, m_i^3, m_i^4) : m_i^1[i] = \theta_i\}.$$

Then, we have $b^{\Theta}(\mathcal{S}) \geq \mathcal{S}$, which implies that $\mathcal{S} \leq \mathcal{S}^{\infty}$.

Proof. Pick any $i \in I$, $\theta_i \in \Theta_i$, and $m_i \in \mathcal{S}_i(\theta_i)$. Fix some $\theta_{-i} \in \Theta_{-i}$ and pick any $\tilde{m}_{-i} \in \mathcal{S}_{-i}(\theta_{-i})$ such that $\tilde{m}_j^1[i] = \theta_i$ and $\tilde{m}_j^1[j] = \theta_j$, for all $j \neq i$. Let the belief $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$ be such that $\psi_i(\theta_{-i}, \tilde{m}_{-i}) = 1$. When individual i of payoff type θ_i holds the belief ψ_i and plays m_i, then she expects the payoff of $u_i\big(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\big)$. On the one hand, if she deviates to $\hat{m}_i$ such that $\hat{m}_i^1[i] = \theta_i^{'}$ and $\hat{m}_i^2 = 1$, then she expects the payoff of $u_i\big(f(\theta_i^{'}, \theta_{-i}), (\theta_i, \theta_{-i})\big)$, which is not improving due to semi-strict EPIC. On the other hand, if she deviates to $\hat{m}_i$ such that $\hat{m}_i^2 > 1$, then she expects the payoff of

$$\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) u_i\big(\hat{m}_i^3[\theta_i](\theta_{-i}), (\theta_i, \theta_{-i})\big) + \left(1 - \frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) u_i\big(y_i^{\theta_i}(\theta_{-i}), (\theta_i, \theta_{-i})\big).$$

As $\hat{m}_i^3[\theta_i] \in Y_i^*[\theta_i]$, she cannot improve by any such deviation. Hence, $m_i \in b_i^{\Theta}(\mathcal{S})[\theta_i]$. This completes the proof of Step 3. □

Step 4: Condition (2) in Theorem 3.1 is satisfied by the constructed mechanism

Proof. Pick $i \in I$, $\theta_i \in \Theta_i$ and $z_i^1 \in Z_i^1$. For each $\theta_{-i} \in \Theta_{-i}$, pick some $\tilde{m}_{-i} \in M_{-i}$ such that $\tilde{m}_j^1[i] = \theta_i$, $\tilde{m}_j^1[j] = \theta_j$, and $\tilde{m}_j^2 = 1$ for all $j \neq i$. From Step 3, it follows that $\tilde{m}_{-i} \in \mathcal{S}_{-i}^{\infty}(\theta_{-i})$. Define the belief $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$ such that $\psi_i(\theta_{-i}, \tilde{m}_{-i}) = z_i^1(\theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$.

By construction, $\psi_i(\theta_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in \mathcal{S}_{-i}^{\infty}(\theta_{-i})$ and $\text{marg}_{\Theta_{-i}}\psi_i = z_i^1$. When individual i of payoff type θ_i holds the belief ψ_i and plays $m_i = (m_i^1, 1, m_i^3, m_i^4)$ such that $m_i^1[i] = \theta_i$, then she expects the payoff of $\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\big)$. On the one hand, if she deviates to $\hat{m}_i$ such that $\hat{m}_i^1[i] = \theta_i^{'}$ and $\hat{m}_i^2 = 1$, then she expects the payoff of $\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(f(\theta_i^{'}, \theta_{-i}), (\theta_i, \theta_{-i})\big)$, which is not improving due to semi-strict EPIC. On the other hand, if she deviates to $\hat{m}_i$ such that $\hat{m}_i^2 > 1$, then she expects the payoff of

$$\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(\hat{m}_i^3[\theta_i](\theta_{-i}), (\theta_i, \theta_{-i})\big) + \left(1 - \frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(y_i^{\theta_i}(\theta_{-i}), (\theta_i, \theta_{-i})\big).$$

As $\hat{m}_i^3[\theta_i] \in Y_i^*[\theta_i]$, she cannot improve by any such deviation. Hence,

$$\arg\max_{m_i^{'} \in M_i} \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(m_i^{'}, m_{-i}), (\theta_i, \theta_{-i})\big) \neq \emptyset,$$

which completes the proof of Step 4. □

Steps 1 through 4 complete the proof of Theorem 5.5. □

References

[1] Aliprantis, C.D., and K.C. Border, "Infinite Dimensional Analysis: A Hitchhiker's Guide," *Springer-Verlag*, (2006).

[2] Bergemann, D., and S. Morris, "Robust Mechanism Design," *Econometrica*, 73, (2005), 1771-1813.

[3] Bergemann, D., and S. Morris, "Strategic Distinguishability with an Application to Robust Virtual Implementation," *Working Paper*, (2007).

[4] Bergemann, D., and S. Morris, "Robust Implementation in Direct Mechanisms," *Review of Economic Studies*, vol. 76, (2009), 1175-1206.

[5] Bergemann, D., and S. Morris, "Robust Implementation in General Mechanisms," *Working Paper*, (2010).

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

[7] Bergemann, D., S. Morris, and S. Takahashi, "Interdependent Preferences and Strategic Distinguishability," *Journal of Economic Theory*, vol. 168, (2017), 329-371.

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

[9] Blackwell, D., and M.A. Girshick, "*Theory of Games and Statistical Decisions*," John Wiley & Sons, Inc., 1954.

[10] Chen, Y-C, T. Kunimoto, Y. Sun, and S. Xiong, "Maskin Meets Abreu and Matsushima," *mimeo*, 2019.

[11] Chen, Y-C, T. Kunimoto, Y. Sun, and S. Xiong, "Rationalizable Implementation in Finite Mechanisms," *mimeo*, 2020.

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

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

[14] Jain, R., "Rationalizable Implementation of Social Choice Correspondences," *Games and Economic Behavior*, vol. 127, (2021), 47-66.

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

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

[17] Xiong, S., "Rationalizable Implementation I: Social Choice Functions," *Working Paper*, (2018).