3-2024

Robust implementation in rationalizable strategies in general mechanisms

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

Rene SARAN

Robust Implementation in Rationalizable Strategies in General Mechanisms*

Takashi Kunimoto† Rene Saran‡

This Version: March 2024

Abstract

A social choice function (SCF) is robustly implementable in rationalizable strategies (RoRat-implementable) if every (interim correlated) rationalizable outcome on every type space agrees with the SCF. We establish that RoRat-implementation is equivalent to weak rationalizable implementation, an implementation notion based on belief-free rationalizability. Applying this equivalence, we identify weak robust monotonicity (weak RM) as the characterizing condition for RoRat-implementation. We show that weak RM is equivalent to semi-strict ex post incentive compatibility and the preference-reversal condition. Furthermore, we clarify the relationships between different "robust" and "rationalizable" implementation notions discussed in the literature. In particular, we prove that strict robust monotonicity (strict RM) characterizes robust implementation in interim equilibria (RoEq-implementation), closing a gap in the literature. We present an example in which weak RM is strictly weaker than strict RM. Thus, RoRat-implementation may be more permissive than RoEq-implementation. We apply our results to quasilinear environments and provide a comprehensive discussion on additional implications of RoRat-implementation.

JEL: C72; D78; D80

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

*This paper subsumes Kunimoto and Saran (2020). We are grateful to Pierpaolo Battigalli, Andrés Carvajal, Antonio Penta, Roberto Serrano, and Takuro Yamashita for helpful comments. We thank the audience at various conferences and seminars for useful comments. This research is supported by the Ministry of Education, Singapore under MOE Academic Research Fund Tier 2 (MOE-T2EP402A20-0007).

†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 agree with the SCF.

A type space defines each player's beliefs and higher-order beliefs about other player's payoff types. Standard Bayesian implementation settings (e.g., Jackson 1991) assume that a specific type space is common knowledge among the players. The designer is therefore able to exploit the players' beliefs in the construction of the implementing mechanism. Naturally, whether an SCF can and cannot be implemented in Bayesian settings crucially depends on the specification of the assumed type space. Following the seminal work of Bergemann and Morris (henceforth, BM, 2005b, 2009a, 2009b, 2011), we take a global approach to robustness by requiring that the mechanism implement the SCF for all possible type spaces which are consistent with the underlying payoff environment.[1]

The solution concept adopted by the designer encapsulates a theory of how players behave in strategic settings. We study the implication of (interim correlated) rationalizability as the solution concept. Rationalizability has some remarkable properties in that regard. It is a set-valued concept that, in contrast to equilibrium, does not require players to have correct conjectures about others' strategies. Furthermore, a type's rationalizable strategies depend only on its belief hierarchy and not on "redundant" features of the type space (Dekel et al., 2007). On a given type space, rationalizability characterizes outcomes that are consistent with common certainty of rationality and the type space (Dekel et al., 2007). Finally, rationalizability also characterizes outcomes that are robust to players gaining information by observing payoff-irrelevant signals (BM, 2017).

The union of rationalizable strategies on all type spaces is characterized by the *belief-free rationalizability* correspondence. The latter is defined by an iterative elimination procedure such as the one for rationalizability, but now on the domain of *payoff* types. Nevertheless, RoRat-implementation is not the same as implementation in belief-free rationalizability (henceforth, BfRat-implementation), as demonstrated by the example in Appendix 11.1. Indeed, our first main result is that RoRat-implementation is equivalent to weak rationalizable implementation (henceforth, wRat-implementation). Introduced in the appendix in

[1]BM (2009a, 2011) and Ollár and Penta (2017) are closely related to our work, as further discussed in the paper. Unlike us, BM (2005b) focus on robust partial implementation whereas BM (2009b) study robust virtual or approximate implementation in finite mechanisms. Artemov et al. (2013) examine robust virtual implementation using the solution concept of Δ-rationalizability, which allows for general belief restrictions. Guo and Yannelis (2022) study robust coalitional implementation where the designer also takes into account the possibility of collusion by a coalition of players.

BM (2010), wRat-implementation is BfRat-implementation plus an extra restriction that the best responses to all first-order beliefs exist. The extra restriction is needed to guarantee that rationalizable strategies are nonempty on all type spaces. The equivalence between RoRat-implementation and wRat-implementation allows us to utilize the latter as an instrument to characterize the former. This approach has the advantage that we do not need to consider rationalizable strategies in different type spaces; instead, as wRat-implementation is defined by imposing conditions on the belief-free rationalizability correspondence, we can simply focus on that correspondence.

In our second main result, we identify *weak robust monotonicity* (weak RM) as the key condition that characterizes RoRat-implementable SCFs under a mild restriction on the environment. Although weak RM could be difficult to check directly, its formulation provides a clean comparison with another important condition in the literature, as further discussed below. Still, in order to provide more insight into the condition, we show that weak RM comprises of incentive and monotonicity-type constraints that are typically found in implementation theory. Specifically, in private-value environments, weak RM is equivalent to *semi-strict* ex post incentive compatibility (semi-strict EPIC). In interdependent-value environments, weak RM is equivalent to semi-strict EPIC and the *preference-reversal* condition. The preference-reversal condition is a monotonicity-type restriction on players' *ex post* preferences. It is equivalent to semi-strict EPIC in private-value environments but not more generally.

As an application of our results, we consider quasilinear environments with monetary transfers. In these environments, a deterministic and interior SCF is RoRat-implementable if it satisfies semi-strict EPIC and the *sign-preserving* property. In fact, when all agents are risk neutral, semi-strict EPIC and the sign-preserving property characterize all (stochastic or deterministic) RoRat-implementable SCFs that have interior transfers. The sign-preserving property is even easier to check than the preference-reversal condition, and is intuitively appealing. It compares the effect on an agent's marginal valuation at her *expected* allocation when she is the sole liar to when everyone else also lies. The property requires that, for at least one agent, either the two effects are in the same direction (positive or negative), or the effect when she is the sole liar dominates the effect when everyone else also lies.

This paper complements the work on RoRat-implementation by BM (2009a) and Ollár and Penta (2017). In contrast to our discrete setting, both papers assume that each player's payoff type is a continuous variable that forms a compact subset of the real line. In single crossing aggregator (SCA) environments, BM (2009a) prove that, for *responsive* SCFs, *strict* ex post incentive compatibility (strict EPIC) and the *contraction* property characterize RoRat-implementation by mechanisms with a *compact* message space. Indeed, they

show that such an SCF can be RoRat-implemented using the direct mechanism. The contraction property requires that the degree of preference interdependence is sufficiently low, which guarantees that truthful revelation in the unique rationalizable strategy. Ollár and Penta (2017) also focus on direct mechanisms and responsive SCFs but allow for general belief restrictions in environments with monetary transfers. Their key insight is that the (im)possibility of RoRat-implementation hinges on the strength of *strategic externalities* in the mechanism for any given level of preference interdependence. Unlike preference interdependence, strategic externalities can be manipulated by the designer using her information about the players' beliefs. Such manipulation is not feasible in our setting; thus the degree of preference interdependence, reflected in the sign-preserving property in quasilinear environments or the preference-reversal condition more generally, is the key to RoRat-implementation.

The focus on direct mechanisms and responsive SCFs in BM (2009a) and Ollár and Penta (2017) however has its limitations. Example 8.1 presents an (approximately) efficient, RoRat-implementable SCF in an SCA auction environment, which cannot be RoRat-implemented by the direct mechanism. This is true even though the SCF is responsive, and satisfies strict EPIC and the contraction property, highlighting an important gap between discrete and continuous settings. Example 8.3 demonstrates starkly the significance of *non-responsive* SCFs in RoRat-implementation. In a social decision setting, that example shows that none of the responsive SCFs is RoRat-implementable but there exists an (approximately) efficient, non-responsive SCF that is RoRat-implementable.[2]

Another important contribution of this paper is to clarify the relationships between different "robust" and "rationalizable" implementation notions discussed in the literature (see Figure 1 and Footnote 11). The relation between RoRat-implementation and robust implementation in interim equilibria (henceforth, RoEq-implementation) deserves a special mention. BM (2011) show that *strict* robust monotonicity (strict RM) is necessary and almost sufficient for RoEq-implementation. We firstly close this gap between the necessary and sufficient conditions by showing that strict RM in fact characterizes RoEq-implementation under our mild restriction on the environment. Secondly, we show that strict RM implies weak RM

[2]There are several other economically relevant environments where the desired SCF is non-responsive. For instance, consider a voting environment where 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 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 8.6).

but the converse is not true (Example 8.6). Hence, if an SCF is RoEq-implementable, then it is RoRat-implementable but the converse is not true. Thus, robustness considerations do not in general make the difference between rationalizable strategies and equilibria moot.

Like BM (2011), we rely on a countably infinite mechanism with an integer game construction to characterize RoRat-implementation. Such canonical mechanisms are commonly deployed in implementation theory but their use has been criticized (see, for e.g., Jackson, 1992). It is worth emphasizing though that the canonical mechanisms are not necessarily meant to be practical devices; these mechanisms serve a different purpose, that being, to understand the constraints imposed by the solution concept on the set of implementable SCFs. With that purpose in mind, we give the designer maximum flexibility in terms of the mechanisms she can use. A natural next step is to limit the designer to a class of mechanisms (e.g., direct, finite etc.) to understand the additional constraints imposed by that restriction. This two-step approach helps to clearly demarcate the effect of the solution concept from that of the mechanism. Indeed, as we show, there is no difference between BfRat-implementation, RoRat-implementation, and RoEq-implementation if we restrict the designer to mechanisms that satisfy the *nonempty best response* property.

Another strand of the literature examines implementation in rationalizable strategies on a *fixed* type space. In 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). They also give an example of a Nash implementable SCF that is not implementable in rationalizable strategies. Recently, Xiong (2023) has provided a complete characterization of SCFs that are implementable in rationalizable strategies under complete information when there are at least three agents. The implementing mechanism in Xiong (2023) 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.[3] In a subsequent paper, Kunimoto et al. (2023), we show that "weak interim rationalizable monotonicity" is necessary and almost sufficient for implementation in rationalizable strategies in incomplete-information environments. That paper also shows that implementation in rationalizable strategies can be more permissive than Bayesian implementation in some incomplete-information environments, including those with private values. In contrast, RoRat-implementation coincides with RoEq-implementation under private values (see Section 7).

The rest of the paper is organized as follows. We present the preliminary definitions in

[3]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).

Section 2. In Section 3, we connect RoRat-implementation to belief-free rationalizability and show that RoRat-implementation is equivalent to wRat-implementation. In Section 4, we show that weak RM characterizes RoRat-implementation. In Section 5, we show that weak RM is equivalent to semi-strict EPIC and the preference-reversal condition. We apply our results to quasilinear environments in Section 6. We compare RoRat-implementation and RoEq-implementation in Section 7. Section 8 provides three examples to illustrate the significance of indirect mechanisms and non-responsive SCFs for RoRat-implementation, and the connections between RoRat-implementation and other notions of implementation. In Section 9, we provide a comprehensive discussion on additional implications of RoRat-implementation. Section 10 concludes. The Appendix contains the example that demonstrates the gap between BfRat-implementation and wRat-implementation, and 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.[4] A *payoff state* is $\theta \in \Theta \equiv \times_{i \in I} \Theta_i$. Denote $\Theta_{-i} \equiv \Theta_1 \times \cdots \times \Theta_{i-1} \times \Theta_{i+1} \times \cdots \times \Theta_n$.[5]

There is a countable set of alternatives A with at least two elements. We assume that A is a separable metrizable space such that its closure, $\bar{A}$, is compact.

For any set X, we will use $\Delta(X)$ to denote the set of probability measures over X. As A is a separable metrizable space, so is $\Delta(A)$ under the weak* topology (Aliprantis and Border, 2006, Theorem 15.12). Therefore, $\Delta(A)$ contains a countable dense subset, which we denote by $\Delta^*(A)$.

We denote an arbitrary probability measure in $\Delta(\bar{A})$ by ℓ. For any $a \in \bar{A}$, we abuse notation and use a to denote the degenerate probability measure that puts probability 1 on a. We use the term *lottery* for any probability measure in $\Delta(\bar{A})$ with a countable support. Note that, since A is countable, any probability measure in $\Delta(A)$ is a lottery. For any lottery ℓ, let $\ell[a]$ be the probability assigned by ℓ to $a \in \bar{A}$. 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 $\alpha_z \geq 0$, for all $z \in \mathbb{Z}$, 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

[4] BM (2011) allow Θ_i to be countable, for all $i \in I$. But that inadvertently introduces an error in their paper: The canonical mechanism constructed in the proof of their Theorem 2 is not necessarily countable, contradicting their restriction to countable mechanisms. We assume a finite payoff-type space to avoid a similar error.

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

probability α_z.

Preferences of player i are represented by the von Neumann-Morgenstern expected utility function $u_i : \Delta(\bar{A}) \times \Theta \to \Re$. Since a lottery has a countable support, for any payoff state θ and lottery ℓ, we can write $u_i(\ell, \theta) = \sum_{a \in A} \ell[a] u_i(a, \theta)$. It is worth emphasizing that, while the utility function is defined on the domain $\Delta(\bar{A})$, mechanisms, as defined below, can only realize outcomes in $\Delta(A)$.

The environment is one of *private values* if the utility of each player $i \in I$ is independent of the other players' payoff types $\theta_{-i} \in \Theta_{-i}$. If not, then the environment has *interdependent values*. In a private-value environment, we simplify notation to write the expected utility of player i as a function of the lottery and her own payoff type, i.e., $u_i : \Delta(\bar{A}) \times \Theta_i \to \Re$.

For each $i \in I$, let $Z_i^1 = \Delta(\Theta_{-i})$ be the set of all possible *first-order beliefs* (i.e., beliefs about the payoff types of the other players) that player i can have. Throughout, we restrict to environments that satisfy a mild assumption on the players' preferences:

Assumption 2.1 (*No-Complete-Indifference*)**.** For each $i \in I$, $\theta_i \in \Theta_i$, and $z_i^1 \in Z_i^1$, there exist $a, a' \in A$ such that

$$\sum_{\theta_{-i} \in \Theta_{-i}} z_i^1(\theta_{-i}) u_i\big(a, (\theta_i, \theta_{-i})\big) \neq \sum_{\theta_{-i} \in \Theta_{-i}} z_i^1(\theta_{-i}) u_i\big(a', (\theta_i, \theta_{-i})\big).$$

No-complete-indifference rules out indifference across alternatives regardless of a player's belief about the payoff types of the other players. It is trivially satisfied in economic environments with money, a private good that is desirable in greater quantities in all payoff states. BM (2009b), too, assume no-complete-indifference to characterize robust virtual implementation in finite mechanisms. Abreu and Matsushima (1992) and Serrano and Vohra (2005) make analogous assumptions in Bayesian settings. See the discussion in Section 9.3 for environments in which no-complete-indifference is violated.

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 the type profile t_{-i} of the other players. 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.

Given the type space $\mathcal{T}$, for each player $i \in I$ 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 (SCF)* $f : \Theta \to \Delta(A)$. The SCF f is *deterministic* if $f(\theta) \in A$ (more formally, $f(\theta)$ is a degenerate lottery), for all $\theta \in \Theta$.

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 $\theta_i' \sim_i^f \theta_i$. We say that individual i is *relevant* for the SCF f if there exist $\theta_i, \theta_i' \in \Theta_i$ such that f is responsive to θ_i and θ_i'; otherwise, we say i is *irrelevant* for f. Without loss of generality, we assume that all individuals $i \in I$ are relevant for the SCF f.[6]

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 $i \in I$. The mechanism $\Gamma = ((M_i)_{i \in I}, g)$ is the *direct mechanism* if $M_i = \Theta_i$, for all $i \in I$, and $g(\theta) = f(\theta)$, for all $\theta \in \Theta$.

2.3 Interim Correlated Rationalizability

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 all $i \in I$ 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 all $i \in I$ and $t_i \in T_i$.

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

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

[6]The problem is trivial if all individuals are irrelevant for the SCF since then the SCF is constant. In case there are some relevant (I^*) and some irrelevant ($I \backslash I^*$) individuals, then the planner can simply ignore the messages of the irrelevant individuals in the mechanism. For instance, in the canonical mechanism constructed to prove Theorem 4.3, the three rules will be defined in the same manner while conditioning only on the messages of the individuals in I^*. Moreover, since the planner can ignore the irrelevant individuals, all of our results are obtained in environments where the no-complete-indifference condition applies only to individuals in I^*.

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

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 and the type space (Dekel et al., 2007, Proposition 2).

2.4 Robust Implementation in Rationalizable Strategies

We now define robust implementation in rationalizable strategies (RoRat-implementation). To do so, we start by defining what we mean by implementation in rationalizable strategies on a specific type space.

Definition 2.2. 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 2.3. 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.

3 Connections with Belief-Free Rationalizability

In this section, we will show that insisting on implementation in rationalizable strategies that is robust to the underlying type space forces the solution concept to be "belief-free" and depend only on the payoff types of the individuals.

To that end, we first define belief-free rationalizability.[7] 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 $\mathcal{S}_i : \Theta_i \to 2^{M_i}$.

[7]Our definition coincides with the definition of belief-free rationalizability in Bergemann and Morris (2017) for a known payoff-type environment (i.e., each player knows his own payoff type and thinks every payoff-type profile of other players is possible), except that we allow for countable mechanisms.

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:

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

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

$\mathcal{S}^\infty$ is the *belief-free rationalizable* message correspondence profile. Belief-free rationalizability is a special case of Δ-rationalizability of Battigalli and Siniscalchi (2003), where Δ denotes a set of restrictions on the first-order beliefs of the players. (These restrictions affect the best response operator above by constraining the marginal of ψ_i on Θ_{-i} to be consistent with Δ.) For each payoff type of each player, Δ-rationalizability characterizes the messages that are consistent with common certainty of rationality and the belief restriction Δ. In the case of belief-free rationalizability, Δ is unrestricted so that players can hold arbitrary first-order beliefs.

Battigalli and Siniscalchi (2003) show that belief-free rationalizability characterizes interim equilibria on all type spaces. Belief-free rationalizability is similarly related to rationalizable strategies: The former is equivalent to the union of rationalizable strategies over all type spaces. For the sake of completeness, we state and prove this result next.

Lemma 3.1. *Consider any mechanism* Γ *such that* $\mathcal{S}_i^\infty(\theta_i) \neq \emptyset$*, for all* $\theta_i \in \Theta_i$ *and* $i \in I$*. 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. ($\Rightarrow$) Battigalli and Siniscalchi (2003, Proposition 4.3) or 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, $m \in B^\infty(t)$.[8]

[8]The assumption $\mathcal{S}_i^\infty(\theta_i) \neq \emptyset$, for all $\theta_i \in \Theta_i$ and $i \in I$, ensures that the constructed type space $\mathcal{T}$ satisfies the restriction that $\hat{\theta}_i : T_i \to \Theta_i$ is surjective, for all $i \in I$.

($\Leftarrow$) 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. □

The above result confirms that belief-free rationalizability encapsulates the implications of imposing robustness with respect to the type space on rationalizable strategies. Thus, one might be inclined to conjecture that RoRat-implementation is equivalent to "implementation in belief-free rationalizability", by which we mean the following:

Definition 3.2. A mechanism $\Gamma = ((M_i)_{i \in I}, g)$ *implements in belief-free rationalizability (or BfRat-implements) the SCF* f if

1. (uniqueness) $m \in \mathcal{S}^{\infty}(\theta) \Rightarrow g(m) = f(\theta)$; and
2. (nonemptiness) $\mathcal{S}_i^{\infty}(\theta_i) \neq \emptyset$, for all $\theta_i \in \Theta_i$ and $i \in I$.

The SCF f is *implementable in belief-free rationalizability (or, BfRat-implementable)* if there exists a mechanism that BfRat-implements f.[9]

We argue below that the above conjecture is false.

We first show that RoRat-implementation is equivalent to a different implementation notion that too is based on belief-free rationalizability. This equivalence is the key to characterizing RoRat-implementation, as we will show later. In their Appendix, BM (2010) define weak rationalizable implementation (wRat-implementation) as follows:

Definition 3.3. 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.

Notice that BfRat-implementation and wRat-implementation have the same uniqueness requirements. However, the nonemptiness requirement in wRat-implementation implies the nonemptiness requirement in BfRat-implementation. Thus, wRat-implementation is BfRat-implementation with additional restrictions on the existence of best responses for all first-order beliefs.

We now establish that RoRat-implementation is equivalent to wRat-implementation.

Theorem 3.4. *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 prove the necessity part of Theorem 3.4 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

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

[9]This definition is equivalent to the concept of robust exact implementation in Artemov et al. (2013) when Δ is unrestricted. Note that the main focus of Artemov et al. (2013) is robust virtual implementation.

Pick any $\theta \in \Theta$. If $m \in \mathcal{S}^\infty(\theta)$, then it follows from Lemma 3.1 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 3.1 that $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$. Lastly, by construction, $\text{marg}_{\Theta_{-i}} \psi_i = z_i^1$ and

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

We prove the sufficiency part of Theorem 3.4 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 3.1 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. □

The above theorem clarifies the difference between RoRat-implementation and BfRat-implementation. Both implementation notions have identical uniqueness requirements, stemming from the equivalence between belief-free rationalizability and the union of rationalizable strategies over all type spaces. The nonemptiness requirement in BfRat-implementation, however, does not guarantee that rationalizable strategies are nonempty on all type spaces. For that guarantee, we need the stronger nonemptiness requirement of wRat-implementation, as shown in the proof of the theorem. We illustrate this difference in the Appendix by presenting an example of an SCF that is *not* wRat-implementable but BfRat-implementable. Having said that, notice that the two nonemptiness requirements coincide in mechanisms where best responses exist for all beliefs (e.g., finite mechanisms and mechanisms with a compact message space when payoff functions are continuous; see Section 9.4).

4 Characterization: Weak Robust Monotonicity

We now use the equivalence between RoRat-implementation and wRat-implementation (Theorem 3.4) to characterize RoRat-implementation.

A *deception* is a profile of correspondences $\beta = (\beta_1, \ldots, \beta_n)$ such that $\beta_i : \Theta_i \to 2^{\Theta_i}$ 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

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

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 the payoff state (θ_i, θ_{-i}).

Next, define a subset of $Y_i^w[\theta_i]$, as follows:

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

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 the payoff state (θ_i, θ_{-i}). Note that $Y_i[\theta_i] \subseteq Y_i^w[\theta_i]$.

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

$$\sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(y(\theta_{-i}'), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(f(\theta_i', \theta_{-i}'), (\theta_i, \theta_{-i})\big). \qquad (1)$$

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 characterizing RoRat-implementation:

Theorem 4.3. *The SCF f is RoRat-implementable if and only if f satisfies weak RM.*

We relegate the proof of the theorem to the Appendix. Here we give a sketch of the argument for why weak RM is necessary, and present the canonical mechanism used to show that weak RM is sufficient.

To understand the necessity of weak RM, consider an unacceptable deception β. Define the message correspondence profile with payoff-type domain $\mathcal{S}$ such that $\mathcal{S}_i(\theta_i) = \{m_i \in \mathcal{S}_i^\infty(\theta_i') : \theta_i' \in \beta_i(\theta_i)\}$, for all $\theta_i \in \Theta_i$ and $i \in I$. Since β is unacceptable, we cannot have $\mathcal{S} \leq b^\Theta(\mathcal{S})$ because that would imply $\mathcal{S} \leq \mathcal{S}^\infty$. Thus, there must be some payoff type θ_i of some player i, some "imitated" payoff type $\theta_i' \in \beta_i(\theta_i)$, and some message in $\mathcal{S}_i^\infty(\theta_i')$ such that the message is not a best response for the payoff type θ_i to any belief $\hat{\psi}_i \in \Delta(\Theta_{-i} \times M_{-i})$ that is consistent with others playing according to $\mathcal{S}_{-i}$. Due to this consistency requirement, the belief $\hat{\psi}_i$ is in fact a belief about the true θ_{-i}, the "imitated" $\theta_{-i}' \in \beta_{-i}(\theta_{-i})$, and messages $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i}')$.

Fix any belief $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ and $\tilde{\theta}_i \in \Theta_i$. Pick any $\theta_{-i}' \in \Theta_{-i}$ and consider the first-order belief z_i^1 that puts probability 1 on θ_{-i}'. The nonemptiness requirement in wRat-

implementation implies that the payoff type $\tilde{\theta}_i$ of player i has a message in $\mathcal{S}_i^\infty(\tilde{\theta}_i)$ that is a best response to some belief $\psi_i^{(\tilde{\theta}_i,\theta'_{-i})} \in \Delta(\Theta_{-i} \times M_{-i})$ that is consistent with z_i^1 and such that $\psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(\theta'_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in \mathcal{S}_{-i}^\infty(\theta'_{-i})$. By the uniqueness requirement, the best response must result in $f(\tilde{\theta}_i, \theta'_{-i})$ and any deviation can change the outcome only if it is strictly worse for player i in the payoff state $(\tilde{\theta}_i, \theta'_{-i})$.

Now consider the belief $\psi_i^\Gamma \in \Delta(\Theta_{-i} \times M_{-i})$ such that player i assigns probability $\psi_i(\theta_{-i}, \theta'_{-i}) \times \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(\theta'_{-i}, m_{-i})$ to the event that the true payoff profile is θ_{-i}, the "imitated" profile is $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$, and the message $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta'_{-i})$. As argued above, there is some message $m_i \in \mathcal{S}_i^\infty(\theta'_i)$ that is not a best response for the payoff type θ_i to the belief ψ_i^Γ. Due to the uniqueness requirement, when player i plays m_i while holding the belief ψ_i^Γ, then player i believes that she obtains the outcome $f(\theta'_i, \theta'_{-i})$ with probability $\psi_i(\theta_{-i}, \theta'_{-i})$. From the argument in the previous paragraph, it follows that any deviation from m_i results in some $y \in Y_i[\tilde{\theta}_i]$. Thus, the inequality (1) must be satisfied to ensure that m_i is not a best response for the payoff type θ_i to the belief ψ_i^Γ.

To construct the canonical mechanism used to show the sufficiency of weak RM, we first define a countable subset of $Y_i^w[\theta_i]$. Recall that $\Delta^*(A)$ is a countable dense subset of $\Delta(A)$. For each i and θ_i, define

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

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

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:

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

Although the above mechanism shares aspects with standard canonical constructions (e.g., use of the integer game), 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' strategies. This is precisely the kind of flexibility that is needed in order to RoRat-implement an SCF that cannot be robustly implemented in interim equilibria, as discussed in Section 7.

5 Incentive and Preference-Reversal Conditions

While weak RM is the key condition for RoRat-implementation, it is a difficult condition to check directly. In this section, we characterize weak RM in terms of simpler and easier-to-check incentive and preference-reversal conditions, which parallel the incentive and monotonicity conditions that are typically found in the full-implementation literature.

We begin by defining the relevant incentive constraints.

Definition 5.1. 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 compatibility* (semi-strict EPIC) if the above inequality becomes strict whenever $\theta_i \not\sim_i^f \theta_i'$. Finally, the SCF f satisfies *strict ex post incentive compatibility* (strict EPIC) if the above inequality becomes strict whenever $\theta_i \neq \theta_i'$.

While semi-strict EPIC is in general weaker than strict EPIC, the two conditions are equivalent for responsive SCFs.

The next result shows that weak RM implies semi-strict EPIC.

Lemma 5.2. *If the SCF f satisfies weak RM, then it satisfies semi-strict EPIC.*[10]

Proof. 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

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

[10] Here is an indirect proof of this lemma. BM (2010, Lemma 6) show that if f is wRat-implementable, then it satisfies semi-strict EPIC. The lemma follows since weak RM implies wRat-implementation (Theorems 3.4 and 4.3). BM (2011, Lemma 1) show that "robust monotonicity" implies semi-strict EPIC. The above lemma, however, does not follow from BM's result because robust monotonicity is stronger than weak RM (see Remark 7.7).

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

To characterize weak RM, it is helpful to distinguish between private- and interdependent-value environments.

5.1 Private Values

In private-value environments, weak RM is equivalent to semi-strict EPIC.

Proposition 5.3. *Suppose we have a private-value environment. The SCF f satisfies weak RM if and only if f satisfies semi-strict EPIC.*

Proof. Lemma 5.2 shows that weak RM implies semi-strict EPIC. To argue the converse, suppose the SCF f satisfies semi-strict EPIC. Pick any unacceptable deception β. Then there exist $i \in I$, $\theta_i \in \Theta_i$, and $\theta_i' \in \beta_i(\theta_i)$ satisfying $\theta_i' \not\sim_i^f \theta_i$. Fix $\tilde{\theta}_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ satisfying $\psi_i(\theta_{-i}, \theta_{-i}') > 0 \Rightarrow \theta_{-i}' \in \beta_{-i}(\theta_{-i})$.

Define $y : \Theta_{-i} \to \Delta(A)$ as $y(\theta_{-i}) = f(\theta_i, \theta_{-i})$, for all $\theta_{-i} \in \Theta_{-i}$. On the one hand, if $\tilde{\theta}_i \sim_i^f \theta_i$, then $f(\theta_i, \theta_{-i}) = f(\tilde{\theta}_i, \theta_{-i})$, for all $\theta_{-i} \in \Theta_{-i}$. Hence, $y \in Y_i[\tilde{\theta}_i]$. On the other, if $\tilde{\theta}_i \not\sim_i^f \theta_i$, then semi-strict EPIC implies that $u_i\big(f(\tilde{\theta}_i, \theta_{-i}), \tilde{\theta}_i\big) > u_i\big(f(\theta_i, \theta_{-i}), \tilde{\theta}_i\big)$, for all $\theta_{-i} \in \Theta_{-i}$. Hence, again, $y \in Y_i[\tilde{\theta}_i]$. Furthermore,

$$\begin{aligned}\sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(y(\theta_{-i}'), \theta_i\big) &= \sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(f(\theta_i, \theta_{-i}'), \theta_i\big) \\ &> \sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(f(\theta_i', \theta_{-i}'), \theta_i\big),\end{aligned}$$

where the inequality is due to semi-strict EPIC and $\theta_i' \not\succsim_i^f \theta_i$. Hence, f satisfies weak RM. □

Thus, semi-strict EPIC characterizes RoRat-implementable SCFs in private-value environments.

Corollary 5.4. *Suppose we have a private-value environment. The SCF f is RoRat-implementable if and only if f satisfies semi-strict EPIC.*

In fact, under private values, RoRat-implementation can be achieved using the direct mechanism, as noted in the next proposition. This is because, when the SCF satisfies semi-strict EPIC, reporting one's true or equivalent payoff type strictly dominates reporting any other payoff type in the direct mechanism.

Proposition 5.5. *Suppose we have a private-value environment. The SCF f is RoRat-implementable if and only if f is RoRat-implementable by the direct mechanism.*

Proof. If f is RoRat-implementable by the direct mechanism, then it is obviously RoRat-implementable. To argue the converse, suppose f is RoRat-implementable. Then f satisfies semi-strict EPIC. Consider the direct mechanism. Fix a type space $\mathcal{T}$ and pick any type $t_i \in T_i$ of any player $i \in I$. Due to semi-strict EPIC, $u_i\big(f(\hat{\theta}_i(t_i), \theta_{-i}), \hat{\theta}_i(t_i)\big) \geq u_i\big(f(\theta_i', \theta_{-i}), \hat{\theta}_i(t_i)\big)$, for all $\theta_i' \in \Theta_i$ and $\theta_{-i} \in \Theta_{-i}$, with a strict inequality if $\hat{\theta}_i(t_i) \not\succsim_i^f \theta_i'$. Therefore, for type t_i of player i, reporting any $\theta_i \sim_i^f \hat{\theta}_i(t_i)$ is strictly better than reporting any $\theta_i' \not\succsim_i^f \hat{\theta}_i(t_i)$, regardless of the strategies of the other players. 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. □

5.2 Interdependent Values

In interdependent-value environments, weak RM is strictly stronger than semi-strict EPIC. For instance, in Example 8.3, there exist responsive SCFs satisfying semi-strict EPIC that are not RoRat-implementable, and hence, do not satisfy weak RM. We now identify the additional restriction imposed by weak RM in interdependent-value environments.

Definition 5.6. The SCF f satisfies the *preference-reversal condition* if, for all unacceptable deceptions β, 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 $\tilde{\theta}_i \sim_i^f \theta_i'$, there exists $y \in Y_i^w[\tilde{\theta}_i]$ such that

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

for all $\theta_{-i} \in \Theta_{-i}$ and $\theta_{-i}' \in \beta_{-i}(\theta_{-i})$.

The preference-reversal condition is implied by semi-strict EPIC in private-value environments. But the two conditions are independent in interdependent-value environments.

Here is the main result of this section: In interdependent-value environments, semi-strict EPIC and the preference-reversal condition together equal weak RM. Formally:

Proposition 5.7. *Suppose we have an interdependent-value environment. The SCF f satisfies weak RM if and only if f satisfies semi-strict EPIC and the preference-reversal condition.*

Hence, semi-strict EPIC and the preference-reversal condition characterize RoRat-implementable SCFs in interdependent-value environments.

Corollary 5.8. *Suppose we have an interdependent-value environment. The SCF f is RoRat-implementable if and only if f satisfies semi-strict EPIC and the preference-reversal condition.*

6 Application: Quasi-linear Environments

Let $X \equiv \{0, \ldots, x\} \subset \Re_+$ be a countable set such that $x > 0$. An *allocation* is an n-tuple $q \in X^n$. The set of *feasible allocations* is some $Q \subseteq X^n$. Given $q \in Q$, we let q_i denote the allocation of agent $i \in I$. We allow for monetary transfers (positive or negative) from the agents to a planner such that each agent's monetary transfer τ_i is a rational number bounded by some $z > 0$. We define $\tau = (\tau_i)_{i \in I}$.

Let $A = \left\{(q, \tau) \in Q \times \mathbb{Q}^n : |\tau_i| \leq z, \forall i \in I\right\}$ be the set of alternatives, where $\mathbb{Q}$ is the set of rational numbers. Thus, A is separable and, as it is bounded, $\bar{A}$ is compact in the Euclidean topology. Note that $\bar{A} = \bar{Q} \times [-z, z]^n$, where $\bar{Q}$ is the closure of Q.

We assume an interdependent-value environment with quasilinear preferences. Specifically, we assume that there is a *valuation function* $v_i : \Re \times \Theta \to \Re$ such that $u_i\big((q, \tau), \theta\big) = v_i(q_i, \theta) - \tau_i$, for all $(q, \tau) \in \bar{A}$ and $\theta \in \Theta$. We further assume that $v_i(q_i, \theta)$ is differentiable in q_i, for all $\theta \in \Theta$ and $i \in I$. Due to the presence of monetary transfers, the environment satisfies no-complete-indifference.

We therefore have the following corollary:

Corollary 6.1. *[**Quasilinear Environments**] The SCF f is RoRat-implementable if and only if f satisfies semi-strict EPIC and the preference-reversal condition.*

We now present some prominent examples where we can apply the above result to determine RoRat-implementable SCFs:

Bilateral trading: There are a buyer (b) and seller (s) of an indivisible good. Let $X = \{0, 1\}$. Let $Q = \{q \in X^2 : q_b + q_s = 1\}$ be the set of feasible allocations. The interpretation

is that if $(q_b, q_s) = (1, 0)$, then the good is traded whereas if $(q_b, q_s) = (0, 1)$, then the good is not traded. The buyer's value and the seller's cost in the payoff state θ are, respectively, $v(\theta)$ and $c(\theta)$. Let $v_b(q_b, \theta) = v(\theta)q_b$ and $v_s(q_s, \theta) = c(\theta)(q_s - 1)$. Then, if the alternative (q, τ) is implemented in the payoff state θ, the buyer receives a utility of $v_b(q_b, \theta) - \tau_b = v(\theta)q_b - \tau_b$ and the seller receives a utility of $v_s(q_s, \theta) - \tau_s = c(\theta)(q_s - 1) - \tau_s$. $\diamond$

Auction: There are $n \geq 2$ agents who want to obtain an indivisible good from an auctioneer. Let $X = \{0, 1\}$. Let $Q = \{q \in X^n : \sum_{i \in I} q_i \leq 1\}$ be the set of feasible allocations. The interpretation is that if $q \in Q$ is such that $q_i = 1$ and $q_j = 0$, for all $j \neq i$, then the good is allocated to agent i; whereas if $q_i = 0$, for all $i \in I$, then the good is retained by the auctioneer. Each agent i's value for the good is given by $v_i(\theta)$ in the payoff state θ. Let $v_i(q_i, \theta) = v_i(\theta)q_i$, for all $i \in I$. Then, if the alternative (q, τ) is implemented in the payoff state θ, agent i receives a utility of $v_i(q_i, \theta) - \tau_i = v_i(\theta)q_i - \tau_i$. $\diamond$

Social decision: There are $n \geq 2$ agents who face a social decision. The cost of implementing the social decision is $c(\theta)$ in the payoff state θ. Let $X = \{0, 1\}$. Let $Q = \{q \in X^n : q_i = q_j, \forall i \neq j\}$ be the set of feasible allocations. The interpretation is that if $q \in Q$ is such that $q_i = 1$ for all $i \in I$, then the social decision is implemented; whereas if $q_i = 0$, for all $i \in I$, then the social decision is not implemented. Each agent i obtains a value of $v_i(\theta)$, if the social decision is implemented, and zero, otherwise, in the payoff state θ. Let $v_i(q_i, \theta) = v_i(\theta)q_i$, for all $i \in I$. Then, if the alternative (q, τ) is implemented in the payoff state θ, agent i receives a utility of $v_i(q_i, \theta) - \tau_i = v_i(\theta)q_i - \tau_i$. $\diamond$

Allocation of a private good: There are $n \geq 2$. A total of x units of a divisible private good is available, where $x > 0$ is a rational number. Each agent can be allocated any quantity in $X = [0, x] \cap \mathbb{Q}$. In this case, let $Q = \{q \in X^n : \sum_{i \in I} q_i = x\}$ be the set of feasible allocations. Each agent i's value for $q_i \in \Re_+$ units of the private good is given by $v_i(q_i, \theta)$ in the payoff state θ. Then, if the alternative (q, τ) is implemented in the payoff state θ, agent i receives a utility of $v_i(q_i, \theta) - \tau_i$. $\diamond$

Public good's provision: There are $n \geq 2$. A public good can be provided in any quantity in $X = [0, x] \cap \mathbb{Q}$, where $x > 0$ is a rational number. The cost of providing $\hat{x}$ units of the public good is $c(\hat{x}, \theta)$ in the payoff state θ. In this case, let $Q = \{q \in X^n : q_i = q_j, \forall i \neq j\}$ be the set of feasible allocations. Each agent i's value for $q_i \in \Re_+$ units of the public good is given by $v_i(q_i, \theta)$ in the payoff state θ. Then, if the alternative (q, τ) is implemented in the payoff state θ, agent i receives a utility of $v_i(q_i, \theta) - \tau_i$. $\diamond$

6.1 Sign-Preserving Property

For any $\ell \in \Delta(\bar{A})$, let $\left(q^{\ell}, \tau^{\ell}\right)$ denote the expected value of ℓ. Thus, for all $i \in I$, q_i^{ℓ} is the expected allocation and τ_i^{ℓ} is the expected monetary transfer of agent i as per the probability measure ℓ.

Given the SCF f, for all $i \in I$ and $\theta, \theta' \in \Theta$, define

$$w_i(\theta', \theta) \equiv \frac{\partial v_i\left(q_i^{f(\theta')}, \theta\right)}{\partial q_i}.$$

Intuitively, $w_i(\theta', \theta)$ is agent i's marginal valuation at her *expected* allocation when all agents report their payoff types as θ' in the payoff state θ.

The following property plays an important role in what follows:

Definition 6.2. The SCF f satisfies the *sign-preserving property* if, for all unacceptable deceptions β, 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 $\tilde{\theta}_i \sim_i^f \theta_i'$

$$\begin{aligned} &\text{sign}\left(w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big) - w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)\right) \\ &= \text{sign}\left(w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big) - w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)\right) \\ &\neq 0, \end{aligned} \tag{2}$$

for all $\theta_{-i} \in \Theta_{-i}$ and $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

To understand (2), suppose all agents report their payoff types as $(\tilde{\theta}_i, \theta'_{-i})$ when the true payoff state is (θ_i, θ_{-i}). At the implemented outcome, $f(\tilde{\theta}_i, \theta'_{-i})$, agent i's marginal valuation at her expected allocation is $w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big)$. If, instead, the reported payoff types were truthful, then agent i's marginal valuation at her expected allocation would be $w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)$. Thus, $w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big) - w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)$ is the impact on agent i's marginal valuation at her expected allocation when all agents jointly lie in their reports. This effect can be decomposed as

$$\Big(w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big) - w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big)\Big) + \Big(w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big) - w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)\Big).$$

The second term is the impact on agent i's marginal valuation at her expected allocation if she is the sole liar. The first term is the additional impact on agent i's marginal valuation at her expected allocation because other agents are lying too. Equation (2) says that either these two effects are in the same direction (positive or negative) or the impact stemming from the agent's lie dominates the additional impact stemming from the other agents' lies.

6.2 Risk-Neutral Preferences

We say that the *agents are risk neutral* if all agents are risk neutral in all payoff states. In this case, $v_i(q_i, \theta)$ is linear in q_i, and hence, it can be expressed as $v_i(\theta)q_i + \kappa_i(\theta)$. Notice that the agents are risk neutral in the bilateral trading, auction, and social decision examples.

If the agents are risk neutral, we have $\partial v_i(q_i, \theta)/\partial q_i = v_i(\theta)$, for all $\theta \in \Theta$ and $i \in I$. Therefore, the following equivalence is obvious (proof is omitted):

Lemma 6.3. *Suppose the agents are risk neutral. The SCF f satisfies the sign-preserving property if and only if, for all unacceptable deceptions β, 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 $\tilde{\theta}_i \sim_i^f \theta_i'$*

$$\text{sign}\left(v_i(\theta_i, \theta_{-i}') - v_i(\tilde{\theta}_i, \theta_{-i}')\right) = \text{sign}\left(v_i(\theta_i, \theta_{-i}) - v_i(\tilde{\theta}_i, \theta_{-i}')\right) \neq 0, \tag{3}$$

for all $\theta_{-i} \in \Theta_{-i}$ and $\theta_{-i}' \in \beta_{-i}(\theta_{-i})$.

Remark 6.4. If the function $v_i(\theta_i, \theta_{-i})$ is strictly increasing in θ_i, for all $i \in I$, and the SCF f is responsive, then the sign-preserving property is equivalent to the contraction property of BM (2009a). We are not making either of those assumptions here. ◇

We say that the SCF f has *interior transfers* if $\tau_i^{f(\theta)} \in (-z, z)$, for all $\theta \in \Theta$ and $i \in I$. The next result shows that, when the agents are risk neutral, the sign-preserving property is necessary and, if the SCF satisfies semi-strict EPIC and has interior transfers, also sufficient for the preference-reversal condition.

Lemma 6.5. *Suppose the agents are risk neutral. If the SCF f satisfies the preference-reversal condition, then f satisfies the sign-preserving property. The converse is true if f satisfies semi-strict EPIC and has interior transfers.*

We thus have the following corollary: When the agents are risk neutral, semi-strict EPIC and the sign-preserving property characterize RoRat-implementable SCFs with interior transfers.

Corollary 6.6. *Suppose the agents are risk neutral and the SCF f has interior transfers. Then, f is RoRat-implementable if and only if f satisfies semi-strict EPIC and the sign-preserving property.*

6.3 Deterministic SCFs

For each $i \in I$, define the *projection* $\rho_i : \bar{Q} \to [0, x]$ such that $\rho_i(q) = q_i$, for all $q \in \bar{Q}$. We say that $\bar{Q}$ is *rich* if $\rho_i(\bar{Q})$ has a nonempty interior, for all $i \in I$. Notice that $\bar{Q}$ is

rich in the allocation of a private good and public good's provision examples; in both cases, $\rho_i(\bar{Q}) = [0, x]$, for all $i \in I$.

When $\bar{Q}$ is rich, we say that the SCF f has *interior allocations* if $q_i^{f(\theta)}$ is an interior point of $\rho_i(\bar{Q})$, for all $\theta \in \Theta$ and $i \in I$. The SCF f is in the *interior* if f has interior allocations and interior transfers.

The next result shows that, when $\bar{Q}$ is rich and the SCF is deterministic and interior, the sign-preserving property is sufficient for the preference-reversal condition regardless of the agents' risk attitudes.

Lemma 6.7. *Suppose $\bar{Q}$ is rich and the SCF f is deterministic and interior. If f satisfies the sign-preserving property, then it satisfies the preference-reversal condition.*

We thus have the following corollary:

Corollary 6.8. *Suppose $\bar{Q}$ is rich and the SCF f is deterministic and interior. If f satisfies semi-strict EPIC and the sign-preserving property, then f is RoRat-implementable.*

7 Comparison with Robust Implementation in Interim Equilibria

We have assumed the solution concept of (interim correlated) rationaliable strategies, which characterizes behavior consistent with common certainty of rationality and the type space. On any given type space, interim equilibrium is a stronger solution concept than rationalizability. Unlike rationalizability, interim equilibrium assumes that players have correct beliefs about each other's behavior. In this section, we compare our results with that of robust implementation in interim equilibrium (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 7.1. A mechanism $\Gamma = ((M_i)_{i\in I}, g)$ *interim implements the SCF f on the type space $\mathcal{T}$* if (i) (nonemptiness) the game $(\mathcal{T}, \Gamma)$ has an interim equilibrium and (ii) (uniqueness) for every interim equilibrium σ of the game $(\mathcal{T}, \Gamma)$, if $\sigma(t)[m] > 0$, then $g(m) = f(\hat{\theta}(t))$.

Robust implementation in interim equilibria is defined as interim implementation over "all type spaces".

Definition 7.2. 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.

Battigalli and Siniscalchi (2003) show that belief-free rationalizablity is equivalent to the union of interim equilibria on all type spaces. Recall that belief-free rationalizability also characterizes rationalizable strategies on all type spaces (see Lemma 3.1). Thus, if every equilibrium outcome on every type space agrees with the SCF, then so will every rationalizable outcome on every type space, and vice versa. Hence, it might seem that asking for robustness (with respect to the type space) makes the difference between rationalizability and equilibrium moot in the context of implementation theory. However, that intuition overlooks the second condition that must be satisfied in order to achieve robust implementation, viz., the theory of behavior (be it rationalizability or equilibrium) must make a nonempty prediction on *every* type space. That is, irrespective of their beliefs and higher-order beliefs, players must be able to act in accordance with the theory. Due to the greater permissiveness of rationalizability on every type space, the nonemptiness requirement is *a priori* weaker in RoRat-implementation than in RoEq-implementation. In fact, it is strictly weaker, as we argue below.

To build the argument, it is important to recall the main results from BM (2011). BM (2011) analyze RoEq-implementation by connecting it with "rationalizable implementation", which is another implementation notion based on belief-free rationalizability.

Definition 7.3. 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.

Remark 7.4. wRat-implementation and Rat-implementation have the same uniqueness requirement. But the nonemptiness requirement in Rat-implementation is stronger than that in wRat-implementation. Notice the change in the order of quantifiers: "for all θ_i" comes after "there exists a belief ψ_i" in the definition of Rat-implementation. Thus, if an SCF is Rat-implementable, then it is wRat-implementable (or RoRat-implementable). But the converse is not true, as discussed below. ◇

BM (2011, Theorem 3) prove that if a mechanism RoEq-implements an SCF, then the *same* mechanism also Rat-implements the SCF; and the converse is true if the mechanism satisfies the "ex post best response property", which guarantees the nonemptiness of interim equilibria in all type spaces (see Section 9.5).[11]

BM (2011, Theorem 1) identify "strict robust monotonicity" as a necessary condition for Rat-implementation (and hence, for RoEq-implementation too). We present an equivalent definition below.

Definition 7.5. 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

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

[11] Recently, Jain et al. (2023) define the appropriate strengthening of the nonemptiness requirement in Rat-implementation, which results in yet another implementation notion based on belief-free rationalizability – "s-rationalizable implementation" – that they show is equivalent to RoEq-implementation. Using this equivalence, they provide examples to show that if a mechanism Rat-implements an SCF, then the *same* mechanism need not RoEq-implement the SCF. (By BM (2011, Theorem 3), that mechanism must fail the ex post best response property.) Jain et al. (2023) also differentiate between RoEq-implementation (in which the existence requirement can be satisfied by either a pure- or -mixed strategy equilibrium on all type spaces) and RoEq-implementation *in pure-strategy equilibra* (in which a pure-strategy equilibrium must exist on all type spaces). They first argue that RoEq-implementation in pure-strategy equilibria is equivalent to "wr-implementation", which too is based on belief-free rationalizability and was first introduced by Müller (2020) in the context of dynamic mechanisms. Then they provide an example to show that if a mechanism RoEq-implements an SCF, then the *same* mechanism need not RoEq-implement the SCF in pure-strategy equilibria. However, Jain et al. (2023) do not characterize RoEq-implementable SCFs (in pure-strategy equilibria or otherwise). They also do not answer the question whether RoEq-implementation is strictly stronger (in terms of the set of implementable SCFs) than Rat-implementation when there are three or more individuals. (For the case of two individuals, Jain et al. (2023, Corollary 1) show that the set of Rat-implementable and RoEq-implementable SCFs are equal.) We settle these questions in environments satisfying no-complete-indifference; see Theorem 7.9 and Footnote 12.

Definition 7.6. The SCF f satisfies *strict robust monotonicity (strict RM)* if every unacceptable deception β is strictly refutable.

Remark 7.7. 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 $\tilde{\theta}_i$. BM (2011) also define "robust monotonicity", which, if taken at face value, seems weaker than strict RM. However, it can be shown that robust monotonicity and strict RM are in fact equivalent conditions. $\diamond$

BM (2011, Theorem 2 and Corollary 1) show that strict RM and "conditional no total indifference" are sufficient for both RoEq-implementation and Rat-implementation. We provide an equivalent definition of conditional no total indifference below.

Definition 7.8. The SCF f satisfies *conditional no total indifference (NTI)* if, for all $i \in I$, $\theta_i \in \Theta_i$, and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$, there exist $y, y^{'} \in \bigcap_{\tilde{\theta}_i \in \Theta_i} Y_i^w[\tilde{\theta}_i]$ such that

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

Notice that conditional NTI is a property of the SCF whereas we have assumed no-complete-indifference, a property of the agents' preferences. The next result closes the gap between the necessary and sufficient conditions for RoEq-implementation as well as between RoEq-implementation and Rat-implementation in our setting.

Theorem 7.9. *The SCF f is RoEq-implementable $\Leftrightarrow$ f is Rat-implementable $\Leftrightarrow$ f satisfies strict RM.*

Proof. It follows from BM (2011) that the SCF f is RoEq-implementable $\Rightarrow$ f is Rat-implementable $\Rightarrow$ f satisfies strict RM. We complete the argument by showing that f satisfies strict RM $\Rightarrow$ f is RoEq-implementable.

Suppose f satisfies strict RM. Then f satisfies weak RM, and hence, semi-strict EPIC (see Lemma 5.2). The next lemma shows that if f satisfies semi-strict EPIC, then f satisfies conditional NTI.

Lemma 7.10. *If the SCF f satisfies semi-strict EPIC, then f satisfies conditional NTI.*

Proof. Pick $i \in I$, $\theta_i \in \Theta_i$, and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$. For all $\theta^{'}_{-i} \in \Theta_{-i}$, define the lottery

$$\ell^{\theta^{'}_{-i}} = \frac{1}{|\Theta_i|} \sum_{\theta^{'}_i \in \Theta_i} f(\theta^{'}_i, \theta^{'}_{-i}).$$

Since individual i is relevant for the SCF f, for all $\tilde{\theta}_i \in \Theta_i$, there exists $\theta_i' \in \Theta_i$ such that $\theta_i' \ \gamma_i^f \ \tilde{\theta}_i$. Then, as f satisfies semi-strict EPIC, we have $u_i\big(f(\tilde{\theta}_i, \theta_{-i}'), (\tilde{\theta}_i, \theta_{-i}')\big) > u_i\big(\ell^{\theta_{-i}'}, (\tilde{\theta}_i, \theta_{-i}')\big)$, for all $\theta_{-i}' \in \Theta_{-i}$ and $\tilde{\theta}_i \in \Theta_i$.

Let $z_i^1(\theta_{-i}) = \sum_{\theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}')$, for all $\theta_{-i} \in \Theta_{-i}$. By no-complete-indifference, there exist a, a' such that $\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(a, (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(a', (\theta_i, \theta_{-i})\big)$. Pick any $\epsilon \in (0,1)$, and define $y^\epsilon : \Theta_{-i} \to \Delta(A)$ and $y'^\epsilon : \Theta_{-i} \to \Delta(A)$ as follows: $y^\epsilon(\theta_{-i}') = (1-\epsilon)\ell^{\theta_{-i}'} + \epsilon a$ and $y'^\epsilon(\theta_{-i}') = (1-\epsilon)\ell^{\theta_{-i}'} + \epsilon a'$, for all $\theta_{-i}' \in \Theta_{-i}$.

As Θ is finite, we can find sufficiently small but positive ϵ such that

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

for all $\theta_{-i}' \in \Theta_{-i}$ and $\tilde{\theta}_i \in \Theta_i$.

We fix any such small but positive ϵ. Then $y^\epsilon, y'^\epsilon \in \bigcap_{\tilde{\theta}_i \in \Theta_i} Y_i^w[\tilde{\theta}_i]$ and, by construction, $\sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(y(\theta_{-i}'), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta_{-i}'} \psi_i(\theta_{-i}, \theta_{-i}') u_i\big(y'(\theta_{-i}'), (\theta_i, \theta_{-i})\big)$. □

Since f satisfies strict RM and conditional NTI, it follows from BM (2011, Theorem 2 and Corollary 1) that f is RoEq-implementable.[12] □

Thus, strict RM characterizes RoEq-implementable (or Rat-implementable) SCFs. As remarked above, strict RM implies weak RM. Example 8.6 shows that strict RM can be strictly stronger than weak RM. As weak RM characterizes RoRat-implementation, it follows that RoRat-implementation can be strictly weaker than RoEq-implementation (or Rat-implementation).

The gap between RoRat- and RoEq-implementation, however, is only possible for non-responsive SCFs in interdependent-value environments. In private-value environments, an SCF is RoRat-implementable if and only if it is RoRat-implementable by the direct mechanism (see Proposition 5.5). As the direct mechanism satisfies the non-empty best response property, it also RoEq-implements the SCF (see Proposition 9.6). Thus, RoRat-implementation and RoEq-implementation coincide in private-value environments. The same is true for responsive SCFs because strict RM is equivalent to weak RM in that case (see Lemma 9.1).

Figure 1 summarizes the relationships between the different implementation notions discussed in this paper. It is worth reiterating that all these implementation notions impose the same uniqueness requirement on the implementing mechanism; where they differ, if at all, is in the strengths of their respective nonemptiness requirements.

[12] The canonical mechanism in BM (2011) has a pure-strategy interim equilibrium on all type spaces. Thus, the set of RoEq-implementable SCFs is the same regardless of whether we require robust implementation in mixed- or pure-strategy equilibria.

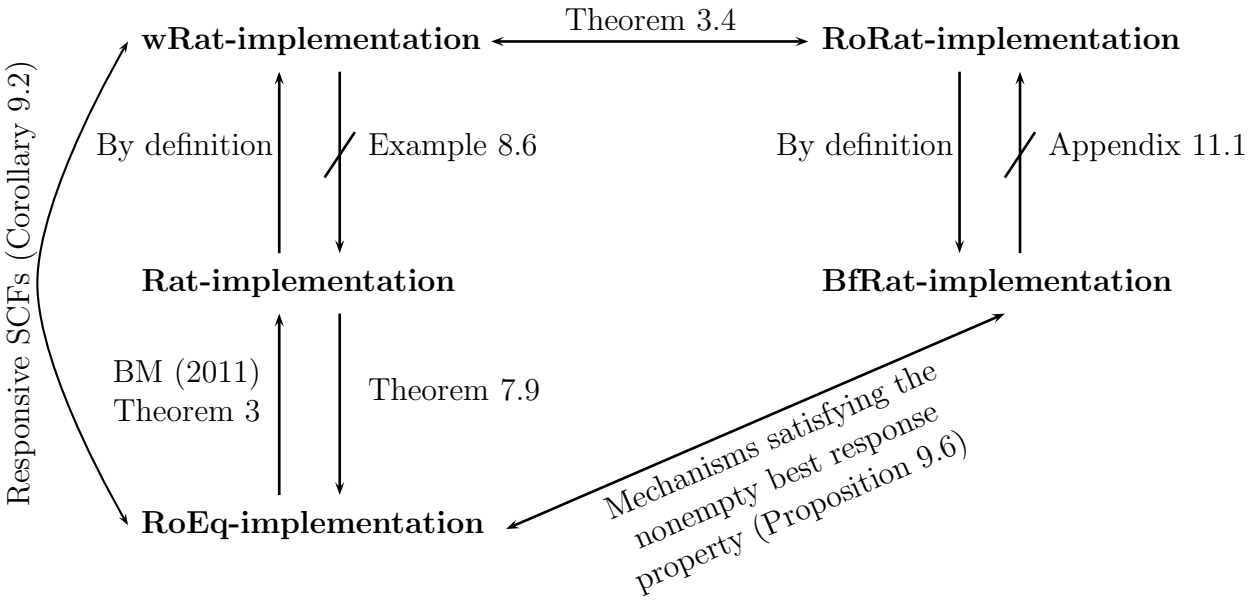

Figure 1: Relationships between different implementation notions.

8 Examples

This section provides three examples. The first two illustrate the significance of indirect mechanisms and non-responsive SCFs for RoRat-implementation, respectively. The third example shows the connections between RoRat-implementation and other notions of implementation discussed in this paper.

Example 8.1 (*Significance of Indirect Mechanisms*)**.** Consider the auction setting discussed in Section 6. Suppose there are two agents and $\Theta_i = \{0, 0.5, 1\}$, for all $i \in I$. Furthermore, suppose the monetary transfers are rational numbers bounded by $z = 2$.

Each agent i's valuation for the object is given by $v_i(\theta_i, \theta_j) \equiv \theta_i + \gamma\theta_j$, where γ is a rational number such that $3/4 \leq \gamma < 1$. Thus, we are in an interdependent-value environment. Notice that the agent's are risk neutral here.

For each $\theta \in \Theta$, we define the set of agents with the highest valuation for the object:

$$W(\theta) = \{i \in I : \ \theta_i + \gamma\theta_{-i} \geq \theta_{-i} + \gamma\theta_i\}.$$

For each $i \in I$ and $\theta \in \Theta$, let $\left(q^i(\theta), \tau^i(\theta)\right) \in A$ denote the alternative in which agent i obtains the object at the price of $(1+\gamma)\theta_j$. Formally, $q_i^i(\theta) = 1$, $q_j^i(\theta) = 0$, $\tau_i^i(\theta) = (1+\gamma)\theta_j$, and $\tau_j^i(\theta) = 0$.

The ex post efficient SCF f^* is defined as follows for all $\theta \in \Theta$: If $W(\theta) = \{i\}$, then $f^*(\theta)$ puts probability 1 on $\left(q^i(\theta), \tau^i(\theta)\right)$ whereas if $W(\theta) = I$, then $f^*(\theta)$ puts equal probability

on $\left(q^i(\theta), \tau^i(\theta)\right)$ and $\left(q^j(\theta), \tau^j(\theta)\right)$. Therefore, for each $i \in I$ and $\theta \in \Theta$, the probability of obtaining the object is

$$q_i^{f^*(\theta)} = \begin{cases} 1/|W(\theta)|, & \text{if } i \in W(\theta), \\ 0, & \text{otherwise,} \end{cases}$$

and the expected monetary transfer is $\tau_i^{f^*(\theta)} = (1+\gamma)\theta_j q_i^{f^*(\theta)}$.

The SCF f^* is responsive and satisfies EPIC but it fails semi-strict EPIC. For instance, $0.5 \; \mathcal{P}_i^{f^*} \; 1$ but the payoff type 0.5 is indifferent between $f^*(0.5, 0)$ and $f^*(1, 0)$ in the payoff state $(0.5, 0)$. Thus, it is impossible to RoRat-implement f^*. However, instead of f^*, the auctioneer can implement an SCF that is arbitrarily close to f^*.

For each $i \in I$, let $\left(\hat{q}^i(\theta), \hat{\tau}^i(\theta)\right) \in A$ denote the alternative in which agent i obtains the object at the price of $(\theta_i + \gamma\theta_j)/2$. Formally, $\hat{q}_i^i(\theta) = 1$, $\hat{q}_j^i(\theta) = 0$, $\hat{\tau}_i^i(\theta) = \theta_i/2 + \gamma\theta_j$, and $\hat{\tau}_j^i(\theta) = 0$.

Consider then the ϵ-efficient allocation rule, as defined in BM (2009a, Section 7): Fix $\epsilon \in (0, 1)$, and for $\theta \in \Theta$, define $f^\epsilon(\theta)$ as follows: The alternative is picked according to $f^*(\theta)$ with probability $(1 - \epsilon)$; for each $i \in I$, the alternative $\left(\hat{q}^i(\theta), \hat{\tau}^i(\theta)\right)$ is picked with probability $\epsilon\theta_i/2$; and the alternative in which the good is kept by the auctioneer and each agent's transfer equals zero is picked with probability $\epsilon \sum_{i \in I}(1 - \theta_i)/2$. Therefore, for each $i \in I$ and $\theta \in \Theta$, the probability of obtaining the object is

$$q_i^{f^\epsilon(\theta)} = \frac{\epsilon}{2}\theta_i + (1 - \epsilon) q_i^{f^*(\theta)}.$$

and the expected monetary transfer is

$$\tau_i^{f^\epsilon(\theta)} = \frac{\epsilon}{4}\theta_i^2 + \frac{\gamma\varepsilon}{2}\theta_j\theta_i + (1 - \epsilon)\tau_i^{f^*(\theta)}.$$

The SCF f^ϵ is responsive and, as argued in BM (2009a, Section 7), satisfies strict EPIC. Assuming that the payoff type space is $[0, 1]$, for all $i \in I$, BM (2009a) show that in this two-bidder auction environment, any responsive SCF that satisfies strict EPIC is RoRat-implementable by the direct mechanism as long as the level of preference interdependence γ is strictly less than 1. The same result does not hold for f^ϵ when the payoff types are finite, as proven in the claim below.

Claim 8.2. *f^ϵ is not RoRat-implementable by the direct mechanism.*

Proof. Suppose f^ϵ is RoRat-implementable by a direct mechanism $\Gamma = ((M_i)_{i \in I}, g)$. Consider the type space such that $T_i = \{t_i, t_i', t_i''\}$, for all $i \in I$. The payoff types of t_i, t_i', and t_i'' are, respectively, 0, 0.5 and 1. Types t_i, t_i', and t_i'' respectively believe that the other agent is of type t_{-i}'', t_{-i}', and t_{-i} with probability one. Since $3/4 \leq \gamma < 1$, the following strategy profile

forms an interim equilibrium on this type space: For all $i \in I$, type t_i reports 1, type t_i' reports 0.5, and finally type t_i'' reports 0. Thus, in particular, reporting 0 is rationalizable for t_2'' whereas reporting 1 is rationalizable for t_1. But then $f^\epsilon(1,0)$ is implemented when types t_1 and t_2'' play these rationalizable actions, contradicting RoRat-implementation. □

The SCF f^ϵ has interior transfers, and satisfies semi-strict EPIC and the sign-preserving property.[13] Hence, f^ϵ is RoRat-implementable (see Corollary 6.6). However, the auctioneer must consider indirect mechanisms if she wishes to RoRat implement f^ϵ. ◇

Example 8.3 (*Significance of Non-responsive SCFs*)**.** Consider the social decision setting discussed in Section 6. Suppose there are two agents and $\Theta_i = \{0.1, 0.2, 0.5, 0.6, 0.9, 1\}$, for all $i \in I$. Here, an agent's payoff types are clustered, so that they could be categorized as being "low", i.e., $\theta_i \in \{0.1, 0.2\}$, "middle", i.e., $\theta_i \in \{0.5, 0.6\}$, or "high", i.e., $\theta_i \in \{0.9, 1\}$. Furthermore, suppose the cost of implementing the social decision $c(\theta) = 1$, for all $\theta \in \Theta$, and the monetary transfers are rational numbers bounded by $z = 2$.

Agent i's valuation function is $v_i(\theta_i, \theta_j) \equiv \sqrt{\theta_i} + \gamma\theta_j^2$, where $\gamma = (1 - \sqrt{0.9})/(1 - 0.9^2) \approx 0.27$. (In fact, as Θ is finite, there exists an $\alpha > 0$ such that the following claims are true whenever $(1 - \sqrt{0.9})/(1 - 0.9^2) \leq \gamma < (1 - \sqrt{0.9})/(1 - 0.9^2) + \alpha$.) As $\gamma > 0$, we are in an interdependent-value environment. Notice that the agent's are risk neutral here.

We first claim that it is impossible to RoRat-implement a responsive SCF in this environment.

Claim 8.4. *If the SCF f is responsive, then f is not RoRat-implementable.*

Proof. Suppose the SCF f is responsive. We argue that f does not satisfy the sign-preserving property. Consider the deception β such that for all $i \in I$, we have $\beta_i(0.9) = \beta_i(1) = \{0.9, 1\}$ and $\beta_i(\theta_i) = \{\theta_i\}$, for all $\theta_i \leq 0.6$. Now pick any $i \in I$, $\theta_i \in \Theta_i$ and $\theta_i' \in \beta_i(\theta_i)$ such that $\theta_i' \neq \theta_i$. Thus, either $\theta_i = 0.9$ and $\theta_i' = 1$ or $\theta_i = 1$ and $\theta_i' = 0.9$.

Consider the case when $\theta_i = 0.9$ and $\theta_i' = 1$. If for agent $j \neq i$, we pick $\theta_j = 1$ and $\theta_j' = 0.9 \in \beta_j(\theta_j)$, then $v_i(\theta_i, \theta_j') - v_i(\theta_i', \theta_j') = \sqrt{0.9} - 1 < 0$. However,

$$v_i(\theta_i, \theta_j) - v_i(\theta_i', \theta_j') = \sqrt{\theta_i} + \gamma\theta_j^2 - (\sqrt{\theta_i'} + \gamma\theta_j'^2) = -(1 - \sqrt{0.9}) + \gamma(1 - 0.9^2) = 0,$$

where the last equality is due to the specification of γ. Likewise, if $\theta_i = 1$ and $\theta_i' = 0.9$, then we can pick $\theta_j = 0.9$ and $\theta_j' = 1 \in \beta_j(0.9)$. Now, $v_i(\theta_i, \theta_j') - v_i(\theta_i', \theta_j') = 1 - \sqrt{0.9} > 0$. However, again, $v_i(\theta_i, \theta_j) - v_i(\theta_i', \theta_j') = 0$.

[13] Since $\gamma < 1$, the SCF f^ϵ satisfies the contraction property in BM (2009a). In this example, the contraction property is equivalent to the sign-preserving property because $v_i(\theta_i, \theta_j)$ is strictly increasing in θ_i and the SCF is responsive.

It follows from Lemma 6.3 that f does not satisfy the sign-preserving property. As a result, f does not satisfy the preference-reversal condition (see Lemma 6.5). Hence, f cannot be RoRat-implemented (see Corollary 6.1). □

Although we cannot implement any responsive SCF in this environment, we now show that there exists an ϵ-efficient, non-responsive SCF that is RoRat-implementable. To simplify the discussion, we define an SCF in its "reduced" form, i.e., in terms of the probability of implementing the social decision $x(\theta) \in [0,1]$ and the expected monetary transfers of both agents, $\big(\tau_1(\theta), \tau_2(\theta)\big) \in [-2,2]^2$, in each payoff state $\theta \in \Theta$.[14]

Ex post efficiency dictates that the probability of implementing the social decision in each $\theta \in \Theta$ must be such that

$$x^*(\theta) = \begin{cases} 1, & \text{if } \sum_{i\in I} \sqrt{\theta_i}\,(1+\gamma\theta_i^{1.5}) \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

For each $i \in I$ and $\theta_j \in \Theta_j$, where $j \neq i$, define $\vartheta_i(\theta_j) \in [0,1]$ such that

$$\sqrt{\vartheta_i(\theta_j)}\big(1+\gamma(\vartheta_i(\theta_j))^{1.5}\big) = 1 - \sqrt{\theta_j}\,\big(1+\gamma\theta_j^{1.5}\big)\,.$$

In this example, $0.2 < \vartheta_i(0.2) < \vartheta_i(0.1) < 0.5$ and $\vartheta_i(\theta_j) < 0.1$, for all $\theta_j \geq 0.5$. Thus, when one agent has any of the high or middle payoff types, then it is ex post efficient to implement the social decision regardless of the other agent's payoff type. However, if both agents have any of the low payoff types, then it is ex post efficient to not implement the social decision.

Define the expected monetary transfers τ^* such that, for each $i \in I$ and $\theta \in \Theta$,

$$\tau_i^*(\theta) = \begin{cases} 1 - \gamma\big(\vartheta_i(\theta_j)\big)^2 - \sqrt{\theta_j}, & \text{if } \theta_i \geq \vartheta_i(\theta_j), \\ 0, & \text{otherwise.} \end{cases}$$

We pair x^* with τ^* to obtain an ex post efficient SCF f^*. It is easy to see that, for each $i \in I$, the SCF f^* is non-responsive to θ_i and θ_i' if and only if either $\{\theta_i, \theta_i'\} \subset \{0.1, 0.2\}$ or $\{\theta_i, \theta_i'\} \subset \{0.5, 0.6, 0.9, 1\}$. While the SCF f^* satisfies EPIC, it does not satisfy semi-strict

[14] For any $\big(x, (\tau_1, \tau_2)\big) \in [0,1] \times [-2,2]^2$, there exists a lottery $\ell \in \Delta(A)$ such that, under ℓ, the social decision is implemented with probability x and each agent i's expected monetary transfer is equal to τ_i. For each $i \in I$, let ς_{τ_i} be an indicator function that takes the value 1 if $\tau_i > 0$ and zero otherwise. Without loss of generality, suppose $|\tau_1| \leq |\tau_2|$. For each $i \in I$, we let $s^i \in \{-2,0,2\}^2$ be the vector of transfers such that if $j < i$, then $s_j^i = 0$ whereas if $j \geq i$, then $s_j^i = 2(2\varsigma_{\tau_j} - 1)$. Then define the lottery ℓ as follows: It assigns probability $x|\tau_1|/2$ to the alternative $((1,1), s^1)$; probability $(1-x)|\tau_1|/2$ to the alternative $((0,0), s^1)$; probability $x(|\tau_2| - |\tau_1|)/2$ to the alternative $((1,1), s^2)$; probability $(1-x)(|\tau_2| - |\tau_1|)/2$ to the alternative $((0,0), s^2)$; probability $x(1 - |\tau_2|/2)$ to the alternative $\big((1,1),(0,0)\big)$; and probability $(1-x)(1-|\tau_2|/2)$ to the alternative $\big((0,0),(0,0)\big)$.

EPIC. For instance, 0.1 $\not\sim_i^{f^*}$ 0.5 but the payoff type 0.1 is indifferent between $f^*(0.1, 0.5)$ and $f^*(0.5, 0.5)$ in the payoff state $(0.1, 0.5)$. Thus, it is impossible to RoRat-implement f^*.

Although the planner cannot implement f^*, there exists an ϵ-efficient non-responsive SCF that is RoRat-implementable. Fix $\epsilon \in (0, 1)$ and define the probability of implementing the social decision in state θ as

$$x^\epsilon(\theta) = \epsilon \left(\sum_{i \in I} \frac{\min\{\theta_i, 0.9\}}{2} \right) + (1 - \epsilon) x^*(\theta).$$

In all states θ, the probability of implementing the social decision differs between $x^\epsilon(\theta)$ and $x^*(\theta)$ by at most ϵ. We pair $x^\epsilon(\theta)$ with the expected monetary transfers τ^ϵ such that

$$\tau_i^\epsilon(\theta) = \frac{\epsilon}{3} \big(\min\{\theta_i, 0.9\} \big)^{1.5} + \frac{\gamma \epsilon}{2} \theta_j^2 \min\{\theta_i, 0.9\} + (1 - \epsilon) \tau_i^*(\theta),$$

for all $\theta \in \Theta$ and $i \in I$, to obtain the SCF f^ϵ.

The SCF f^ϵ has interior transfers because $\tau_i^\epsilon(\theta) \in (-2, 2)$, for all $\theta \in \Theta$ and $i \in I$. Recall that the SCF f^* is non-responsive to payoff types 0.9 and 1, for each $i \in I$. Therefore, by construction, for each $i \in I$, the SCF f^ϵ is non-responsive to payoff types 0.9 and 1 and it is responsive to any other pair of payoff types. Then it is easy to check that f^ϵ satisfies semi-strict EPIC. We now claim that f^ϵ satisfies the sign-preserving property.

Claim 8.5. *The SCF f^ϵ satisfies the sign-preserving property.*

Proof. Pick any unacceptable deception β. Define the set $R \equiv \big\{ (i, \theta_i, \theta_i') \in I \times \Theta_i^2 : \theta_i' \in \beta_i(\theta_i) \big\}$. Consider, without loss of generality, any tuple $(1, \theta_1, \theta_1')$ that solves

$$\max_{(i, \theta_i, \theta_i') \in R} |\theta_i^2 - \theta_i'^2|.$$

As β is unacceptable, it must be that $\theta_1 \neq \theta_1'$.

First, suppose $\theta_1' \leq 0.6$. Then there does not exist any $\tilde{\theta}_1 \neq \theta_1'$ such that $\tilde{\theta}_1 \sim_1^{f^\epsilon} \theta_1'$. Now, pick any $\theta_2 \in \Theta_2$ and $\theta_2' \in \beta_2(\theta_2)$. It then follows that $(2, \theta_2, \theta_2') \in R$. Hence, $|\theta_2^2 - \theta_2'^2| \leq |\theta_1^2 - \theta_1'^2|$. If $\theta_1 - \theta_1' > 0$, then $v_1(\theta_1, \theta_2') - v_1(\theta_1', \theta_2') > 0$ and

$$\begin{aligned} v_1(\theta_1, \theta_2) - v_1(\theta_1', \theta_2') &= \sqrt{\theta_1} + \gamma \theta_2^2 - \left(\sqrt{\theta_1'} + \gamma \theta_2'^2 \right) \\ &\geq \sqrt{\theta_1} - \sqrt{\theta_1'} - \gamma(\theta_1^2 - \theta_1'^2) \\ &= (\sqrt{\theta_1} - \sqrt{\theta_1'}) \left(1 - \frac{(\sqrt{\theta_1} + \sqrt{\theta_1'})(\theta_1 + \theta_1')}{(1 + \sqrt{0.9})(1 + 0.9)} \right) \end{aligned}$$

$$> 0,$$

where the equality follows from $\gamma = (1-\sqrt{0.9})/(1-0.9^2)$ and the strict inequality follows because $\sqrt{\theta_1}+\sqrt{\theta_1'} < 1+\sqrt{0.9}$ and $\theta_1+\theta_1' < 1+0.9$ since $\theta_1' \leq 0.6$. If, instead, $\theta_1-\theta_1' < 0$, then we can similarly show that $v_1(\theta_1,\theta_2') - v_1(\theta_1',\theta_2') < 0$ and $v_1(\theta_1,\theta_2) - v_1(\theta_1',\theta_2') < 0$.

Second, suppose $\theta_1' \geq 0.9$ and $\theta_1 \leq 0.6$. Then pick any $\tilde{\theta}_1 \sim_1^{f^\epsilon} \theta_1'$, $\theta_2 \in \Theta_2$ and $\theta_2' \in \beta_2(\theta_2)$. It then follows that $\tilde{\theta}_1 = 1$ or 0.9 and $(2,\theta_2,\theta_2') \in R$. Hence, $|\theta_2^2-\theta_2'^2| \leq |\theta_1^2-\theta_1'^2|$. As $\tilde{\theta}_1 \sim_1^{f^\epsilon} \theta_1'$, we have $\theta_1 - \tilde{\theta}_1 < 0$. Then $v_1(\theta_1,\theta_2') - v_1(\tilde{\theta}_1,\theta_2') < 0$ and

$$\begin{aligned} v_1(\theta_1,\theta_2) - v_1(\tilde{\theta}_1,\theta_2') &= \sqrt{\theta_1} + \gamma\theta_2^2 - \left(\sqrt{\tilde{\theta}_1} + \gamma\theta_2'^2\right) \\ &\leq \sqrt{\theta_1} - \sqrt{\tilde{\theta}_1} + \gamma|\theta_1^2 - \theta_1'^2| \\ &= \left(\sqrt{\theta_1'} - \sqrt{\theta_1}\right)\left(-\frac{\sqrt{\tilde{\theta}_1}-\sqrt{\theta_1}}{\sqrt{\theta_1'}-\sqrt{\theta_1}} + \frac{(\sqrt{\theta_1'}+\sqrt{\theta_1})(\theta_1'+\theta_1)}{(1+\sqrt{0.9})(1+0.9)}\right), \end{aligned}$$

where the equality follows because $\gamma = (1-\sqrt{0.9})/(1-0.9^2)$ and $|\theta_1^2-\theta_1'^2| = (\theta_1'-\theta_1)(\theta_1'+\theta_1)$. The above expression is clearly negative if $\tilde{\theta}_1 \geq \theta_1'$, which in turn happens in one of the following cases: $\tilde{\theta}_1 = \theta_1' = 0.9$; $\tilde{\theta}_1 = \theta_1' = 1$; and $\tilde{\theta}_1 = 1$ and $\theta_1' = 0.9$. So, we are left with only one possibility $\tilde{\theta}_1 = 0.9$ and $\theta_1' = 1$. Even in that case, it can be shown that the above expression is negative for all values of $\theta_1 \leq 0.6$.

Finally, suppose $\theta_1' \geq 0.9$ and $\theta_1 \geq 0.9$. Then $\theta_1' \sim_1^{f^\epsilon} \theta_1$. As β is unacceptable, there must exist another tuple $(i,\hat{\theta}_i,\hat{\theta}_i') \in R$ such that $\hat{\theta}_i' \not\sim_i^{f^\epsilon} \hat{\theta}_i$. Then pick any $\tilde{\theta}_i \sim_i^{f^\epsilon} \hat{\theta}_i'$, $\theta_j \in \Theta_j$ and $\theta_j' \in \beta_j(\theta_j)$. It then follows that $(j,\theta_j,\theta_j') \in R$. Hence, $|\theta_j^2 - \theta_j'^2| \leq |\theta_1^2 - \theta_1'^2| = 0.19$ (because $\theta_1', \theta_1 \geq 0.9$ and $\theta_1 \neq \theta_1'$). Assume $\hat{\theta}_i > \tilde{\theta}_i$. Then $v_i(\hat{\theta}_i,\theta_j') - v_i(\tilde{\theta}_i,\theta_j') > 0$. Since $\hat{\theta}_i' \not\sim_i^{f^\epsilon} \hat{\theta}_i$, it follows from the transitivity of $\sim_i^{f^\epsilon}$ that $\tilde{\theta}_i \not\sim^{f^\epsilon} \hat{\theta}_i$. So, the smallest possible value of $\sqrt{\hat{\theta}_i} - \sqrt{\tilde{\theta}_i}$ is obtained when $\hat{\theta}_i = 0.6$ and $\tilde{\theta}_i = 0.5$. But then

$$v_i(\hat{\theta}_i,\theta_j) - v_i(\tilde{\theta}_i,\theta_j') = \sqrt{\hat{\theta}_i} + \gamma\theta_j^2 - \left(\sqrt{\tilde{\theta}_i} + \gamma\theta_j'^2\right) \geq \sqrt{0.6} - \sqrt{0.5} - \gamma(0.19) > 0,$$

where the strict inequality follows because $\gamma = (1-\sqrt{0.9})/(1-0.9^2) = (1-\sqrt{0.9})/0.19$. Next, assume $\hat{\theta}_i < \tilde{\theta}_i$. Then $v_i(\hat{\theta}_i,\theta_j') - v_i(\tilde{\theta}_i,\theta_j') < 0$. Since $\hat{\theta}_i' \not\sim_i^{f^\epsilon} \hat{\theta}_i$, it follows from the transitivity of $\sim_i^{f^\epsilon}$ that $\tilde{\theta}_i \not\sim^{f^\epsilon} \hat{\theta}_i$. So, the largest possible value of $\sqrt{\hat{\theta}_i} - \sqrt{\tilde{\theta}_i}$ is obtained when $\hat{\theta}_i = 0.5$ and $\tilde{\theta}_i = 0.6$. But then

$$v_i(\hat{\theta}_i,\theta_j) - v_i(\tilde{\theta}_i,\theta_j') = \sqrt{\hat{\theta}_i} + \gamma\theta_j^2 - \left(\sqrt{\tilde{\theta}_i} + \gamma\theta_j'^2\right) \leq \sqrt{0.5} - \sqrt{0.6} + \gamma(0.19) < 0,$$

where the strict inequality follows because $\gamma = (1-\sqrt{0.9})/(1-0.9^2) = (1-\sqrt{0.9})/0.19$.

We thus conclude that f^ϵ satisfies the sign-preserving property. □

It then follows from Corollary 6.6 that f^ϵ is RoRat-implementable. ◇

Example 8.6 (*RoRat-implementation is strictly weaker than Rat-implementation*)**.** 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''	$2,0$	$5,0$

It is straightforward to check that the environment satisfies no-complete-indifference.

The SCF f selects the alternative that 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 8.7. *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 8.8. *The SCF f satisfies weak RM.*

Proof. Weak RM is equivalent to semi-strict EPIC and the preference-reversal condition (Proposition 5.7). It is straightforward to check that f satisfies semi-strict EPIC. We show that f satisfies the preference-reversal condition.

First, we consider any unacceptable deception β such that either $\theta_1' \in \beta_1(\theta_1)$ or $\theta_1'' \in \beta_1(\theta_1)$. As $\theta_1' \sim_1^f \theta_1''$, in what follows, we consider each possible case of $\tilde{\theta}_1 \in \{\theta_1', \theta_1''\}$.

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

Define $y : \Theta_2 \to \Delta(A)$ to be such that $y(\theta_2) = a$ and $y(\theta_2') = z'$. It is straightforward to confirm that $y \in Y_1^w[\theta_1']$. Moreover, $u_1\big(y(\tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) = 4 > 3 \geq u_1\big(f(\theta_1', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big)$, for all $\tilde{\theta}_2 \in \Theta_2$ and $\tilde{\theta}_2' \in \beta_2(\tilde{\theta}_2)$.

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

Define $y : \Theta_2 \to \Delta(A)$ to be such that $y(\theta_2) = a$ and $y(\theta_2') = \frac{1}{5}c + \frac{4}{5}z$. It is straightforward to confirm that $y \in Y_1^w[\theta_1'']$. Moreover, $u_1\big(y(\tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big) > u_1\big(f(\theta_1'', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big)$, for all $\tilde{\theta}_2 \in \Theta_2$ and $\tilde{\theta}_2' \in \beta_2(\tilde{\theta}_2)$, because

	$(\tilde{\theta}_2, \tilde{\theta}_2')$ such that $\tilde{\theta}_2' \in \beta_2(\tilde{\theta}_2)$			
	(θ_2, θ_2)	(θ_2, θ_2')	(θ_2', θ_2)	(θ_2', θ_2')
$u_1\big(y(\tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)\big)$	4	16/5	4	11/5
$u_1\big(f(\theta_1'', \tilde{\theta}_2'), (\theta_1, \tilde{\theta}_2)$	0	3	3	2

Second, we consider any unacceptable deception β such that $\theta_2' \in \beta_2(\theta_2)$ and $\beta_1(\theta_1) = \{\theta_1\}$. 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 $y \in Y_2^w[\theta_2']$. Moreover, $u_2\big(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2)\big) > u_2\big(f(\tilde{\theta}_1', \theta_2'), (\tilde{\theta}_1, \theta_2)\big)$, for all $\tilde{\theta}_1 \in \Theta_1$ and $\tilde{\theta}_1' \in \beta_1(\tilde{\theta}_1)$, because $\beta_1(\theta_1) = \{\theta_1\}$ and

	$(\tilde{\theta}_1, \tilde{\theta}_1')$ such that $\tilde{\theta}_1' \in \beta_1(\tilde{\theta}_1)$						
	(θ_1, θ_1)	(θ_1', θ_1)	(θ_1', θ_1')	(θ_1', θ_1'')	(θ_1'', θ_1)	(θ_1'', θ_1')	(θ_1'', θ_1'')
$u_2\big(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2)\big)$	1	2	2	2	2	2	2
$u_2\big(f(\tilde{\theta}_1', \theta_2'), (\tilde{\theta}_1, \theta_2)\big)$	0	1	0	0	0	0	0

Third, we consider any unacceptable deception β such that $\theta_2 \in \beta_2(\theta_2')$ and $\beta_1(\theta_1) = \{\theta_1\}$. 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 $y \in Y_2^w[\theta_2]$. Moreover, $u_2\big(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2')\big) > u_2\big(f(\tilde{\theta}_1', \theta_2), (\tilde{\theta}_1, \theta_2')\big)$, for all $\tilde{\theta}_1 \in \Theta_1$ and $\tilde{\theta}_1' \in \beta_1(\tilde{\theta}_1)$, because $\beta_1(\theta_1) = \{\theta_1\}$ and

	$(\tilde{\theta}_1, \tilde{\theta}_1')$ such that $\tilde{\theta}_1' \in \beta_1(\tilde{\theta}_1)$						
	(θ_1, θ_1)	(θ_1', θ_1)	(θ_1', θ_1')	(θ_1', θ_1'')	(θ_1'', θ_1)	(θ_1'', θ_1')	(θ_1'', θ_1'')
$u_2\big(y(\tilde{\theta}_1'), (\tilde{\theta}_1, \theta_2')\big)$	3/2	3/2	3/2	3/2	1/4	1/4	1/4
$u_2\big(f(\tilde{\theta}_1', \theta_2), (\tilde{\theta}_1, \theta_2')\big)$	0	1	0	0	0	0	0

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 we can

use the fact that f satisfies semi-strict EPIC to generate the required preference reversal. For example, suppose $\theta_1 \in \beta_1(\theta_1^{'})$ – similar argument applies to case when $\theta_1 \in \beta_1(\theta_1^{''})$. Then define $y : \Theta_2 \to \Delta(A)$ 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 $y \in Y_1^w[\theta_1]$ and $u_1\big(y(\tilde{\theta}_2^{'}), (\theta_1^{'}, \tilde{\theta}_2)\big) > u_1\big(f(\theta_1, \tilde{\theta}_2^{'}), (\theta_1^{'}, \tilde{\theta}_2)\big)$, for all $\tilde{\theta}_2 \in \Theta_2$ and $\tilde{\theta}_2^{'} \in \beta_2(\tilde{\theta}_2)$.

We thus conclude that f satisfies the preference-reversal condition. □

9 Further Implications of RoRat-Implementation

9.1 Responsive SCFs

We previously argued that RoRat-implementation is strictly weaker than RoEq-implementation (or Rat-implementation). We did so by pointing to Example 8.6, which shows that strict RM – the characterizing condition for RoEq-implementation – is strictly stronger than weak RM – the characterizing condition for RoRat-implementation. Notice that the SCF in Example 8.6 is non-responsive. Is there a gap between RoRat-implementation and RoEq-implementation for the class of responsive SCFs? The answer is no because strict RM and weak RM coincide for responsive SCFs, as we show next.

Lemma 9.1. *Suppose the SCF f is responsive. Then f satisfies strict RM if and only if f satisfies weak RM.*

Proof. 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 (see Lemma 5.2). 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. □

Thus, we obtain the following corollary:

Corollary 9.2. *Suppose the SCF f is responsive. Then,*

$$f \text{ is RoRat-implementable} \Leftrightarrow f \text{ is Rat-implementable} \Leftrightarrow f \text{ is RoEq-implementable.}$$

9.2 Type Spaces in which the Canonical Mechanism has Interim Equilibria

As discussed earlier, RoRat-implementation is strictly weaker than RoEq-implementation because the nonemptiness requirement in the latter is strictly stronger than that in the former. Any mechanism, in particular the canonical mechanism used to prove Theorem 4.3, 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. Although one might find it pathological that the induced game has no equilibria on *some* type space, notice that the mechanism is well-behaved in terms of rationalizability. After all, rationalizable strategies exist on all type spaces, ensuring that the players have a complete theory of how to play the game, without invoking the strong equilibrium requirement that they have correct conjectures about each other's strategies. Furthermore, the lack of equilibria in *some* type space does not preclude the existence of interim equilibria in other type spaces.

For instance, we now establish that our canonical mechanism has nonempty interim equilibria on all type spaces $\mathcal{T} = (T_i, \hat{\theta}_i, \hat{\pi}_i)_{i\in I}$ that are sufficiently large in the following sense: For all $i \in I$, there exists a surjective mapping $\tau^i : T_i \to \Theta$ such that (i) $\tau^i(t_i)_i = \hat{\theta}_i(t_i)$, for all $t_i \in T_i$ and $i \in I$, and (ii) for all $i \in I$, $t_i \in T_i$ and $t_{-i} \in T_{-i}$,

$$\hat{\pi}_i(t_i)[t_{-i}] > 0 \Rightarrow \tau^j(t_j)_i = \hat{\theta}_i(t_i), \forall j \in I\backslash\{i\},$$

where $\tau^j(t_j)_i \in \Theta_i$ denotes the i-th coordinate of $\tau^j(t_j)$, for all $j \in I$. These conditions imply that each individual $i \in I$ has at least as many types in T_i as the number of payoff states in Θ and if types t_i and t_i' have different payoff types, then the supports of their respective beliefs, $\hat{\pi}_i(t_i)$ and $\hat{\pi}_i(t_i')$, do not intersect. A prominent example of such a type space is the complete-information type space, i.e., when individuals have complete information about the realized payoff state.[15]

Suppose the SCF f satisfies weak RM, so that it is RoRat-implementable by the canonical mechanism Γ constructed in the proof of Theorem 4.3. Pick any type space $\mathcal{T}$, as defined above. We now show that the canonical mechanism Γ has a pure-strategy 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 $\sigma_i(t_i) = (\tau^i(t_i), 1, m_i^3, m_i^4)$, for all $t_i \in T_i$ and $i \in I$. We argue that σ is an interim equilibrium of the game $(\mathcal{T}, \Gamma)$.

Pick individual $i \in I$ of type $t_i \in T_i$. If everyone plays the game $(\Gamma, \mathcal{T})$ according to the strategy profile σ, then the outcome is given by Rule 1 and type $t_i \in T_i$ of individual i expects a payoff of

$$\begin{aligned}
&\sum_{t_{-i}\in T_{-i}} \hat{\pi}_i(t_i)[t_{-i}] u_i\big(f(\tau^i(t_i)_i, (\tau^j(t_j)_j)_{j\in I\backslash\{i\}}), (\hat{\theta}_i(t_i), \hat{\theta}_{-i}(t_{-i}))\big) \\
&= \sum_{t_{-i}\in T_{-i}} \hat{\pi}_i(t_i)[t_{-i}] u_i\big(f(\hat{\theta}_i(t_i), \hat{\theta}_{-i}(t_{-i})), (\hat{\theta}_i(t_i), \hat{\theta}_{-i}(t_{-i}))\big),
\end{aligned}$$

where the equality follows because $\tau^j(t_j)_j = \hat{\theta}_j(t_j)$, for all $t_j \in T_j$ and $j \in I$.

On the one hand, if type t_i 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 $\sum_{t_{-i}\in T_{-i}} \hat{\pi}_i(t_i)[t_{-i}] u_i\big(f(\tilde{\theta}_i, \hat{\theta}_{-i}(t_{-i})), (\hat{\theta}_i(t_i), \hat{\theta}_{-i}(t_{-i}))\big)$, which is not improving due to semi-strict EPIC. On the other hand, if type t_i deviates to $\hat{m}_i$ such that $\hat{m}_i^2 > 1$, then Rule 2 is triggered. Since $\hat{\pi}_i(t_i)[t_{-i}] > 0 \Rightarrow \tau^j(t_j)_i = \hat{\theta}_i(t_i), \forall j \in$

[15]The complete-information type space is such that $T_i = \{t_i^\theta : \theta \in \Theta\}$, for all $i \in I$; $\hat{\theta}_i(t_i^\theta) = \theta_i$ and $\hat{\pi}_i(t_i^\theta)\big[(t_j^\theta)_{j\in I\backslash\{i\}}\big] = 1$, for all $t_i^\theta \in T_i$ and $i \in I$.

$I\backslash\{i\}$, and $\tau^j(t_j)_j = \hat{\theta}_j(t_j)$, for all $t_j \in T_j$ and $j \in I \setminus \{i\}$, she expects the payoff of

$$\sum_{t_{-i}\in T_{-i}} \hat{\pi}_i(t_i)[t_{-i}] \left\{ \begin{array}{l} \left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) u_i\big(\hat{m}_i^3[\hat{\theta}_i(t_i)](\hat{\theta}_{-i}(t_{-i})), (\hat{\theta}_i(t_i), \hat{\theta}_{-i}(t_{-i}))\big) \\ \\ + \left(1 - \frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right) u_i\big(y_i^{\hat{\theta}_i(t_i)}(\hat{\theta}_{-i}(t_{-i})), (\hat{\theta}_i(t_i), \hat{\theta}_{-i}(t_{-i}))\big) \end{array} \right\}.$$

As $\hat{m}_i^3[\hat{\theta}_i(t_i)]$ is chosen from $Y_i^*[\hat{\theta}_i(t_i)]$, type t_i cannot improve her payoff by any such deviation. Hence, the message $\sigma_i(t_i)$ is a best response of type t_i against σ_{-i}, which completes the argument that σ is an interim equilibrium of the game $(\mathcal{T}, \Gamma)$.

More generally, take any mechanism that RoRat-implements an SCF. Then, for every type space in which the mechanism has no interim equilibrium, there exists an "expanded" type space in which the mechanism *has* an interim equilibrium. This is because, by the definition of RoRat-implementation, the set of rationalizable strategies is nonempty on every type space. And every rationalizable strategy profile on every type space can be obtained as a pure-strategy interim equilibrium on another type space (see Lemma 3.1 and its proof). This new type space is basically an expansion of the original type space, where individuals now additionally observe *payoff-irrelevant* signals (Bergemann and Morris (2017, Proposition 7)).

9.3 Environments violating No-Complete-Indifference

We have assumed that the environment satisfies the mild restriction of no-complete-indifference. The following results also hold in environments that violate no-complete-indifference: The equivalence between wRat-implementation and RoRat-implementation (Theorem 3.4); the *necessity* of weak RM for RoRat-implementation in Theorem 4.3; the equivalence between weak RM and semi-strict EPIC in private-values environments (Proposition 5.3); and the equivalence between weak RM, on the one hand, and semi-strict EPIC and the preference-reversal condition, on the other (Proposition 5.7).

Which condition(s) on the SCF is(are) sufficient for RoRat-implementation in environments that violate no-complete-indifference? This remains an open question. In fact, the same question is open for RoEq-implementation too! That might seem surprising in light of the result in BM (2011) that strict RM and conditional NTI – two conditions on the SCF – are sufficient for RoEq-implementation. While that result ostensibly applies to all environments, we now argue that the assumption of conditional NTI in fact implies that the environment satisfies no-complete-indifference.

Lemma 9.3. *Suppose there exists an SCF that satisfies conditional NTI. Then the environ-*

ment satisfies no-complete-indifference.

Proof. Pick any $i \in I$, $\theta_i \in \Theta_i$, and belief $z_i^1 \in Z_i^1$. Pick any $\theta_{-i}^{'} \in \Theta_{-i}$. Then define $\psi_i(\theta_{-i}, \theta_{-i}^{'}) = z_i^1(\theta_{-i})$, for all $\theta_{-i} \in \Theta_{-i}$. Since $\theta_{-i}^{'}$ is fixed, by conditional NTI, there exist lotteries $\ell, \ell^{'} \in \Delta(A)$ such that $\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i(\ell, (\theta_i, \theta_{-i})) > \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i(\ell^{'}, (\theta_i, \theta_{-i}))$, which implies no-complete-indifference. □

It is worth noting though that RoRat-implementation (or RoEq-implementation) is not feasible in arbitrary environments. Indeed, as we show next, if an SCF is RoRat-implementable (or RoEq-implementable), then the environment must satisfy the following weakening of no-complete-indifference.

Definition 9.4. The environment satisfies *weak no-complete-indifference* if, for each $i \in I$, $\theta_i \in \Theta_i$, and $z_i^1 \in Z_i^1$, there exist $y : \Theta_{-i} \to \Delta(A)$ and $y^{'} : \Theta_{-i} \to \Delta(A)$ such that

$$\sum_{\theta_{-i} \in \Theta_{-i}} z_i^1(\theta_{-i}) u_i\big(y(\theta_{-i}), (\theta_i, \theta_{-i})\big) \neq \sum_{\theta_{-i} \in \Theta_{-i}} z_i^1(\theta_{-i}) u_i\big(y^{'}(\theta_{-i}), (\theta_i, \theta_{-i})\big).$$

Lemma 9.5. *If the SCF f is RoRat-implementable (or RoEq-implementable), then the environment satisfies weak no-complete-indifference.*

Proof. If f is RoRat-implementable (or RoEq-implementable), then f satisfies semi-strict EPIC. Pick any $i \in I$, $\theta_i \in \Theta_i$, and belief $z_i^1 \in Z_i^1$. Since all individuals are relevant for the SCF, there exists $\theta_i^{'} \not\sim_i^f \theta_i$. Then, by semi-strict EPIC, we have $u_i\big(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\big) > u_i\big(f(\theta_i^{'}, \theta_{-i}), (\theta_i, \theta_{-i})\big)$, for all $\theta_{-i} \in \Theta_{-i}$. Let $y : \Theta_{-i} \to \Delta(A)$ and $y^{'} : \Theta_{-i} \to \Delta(A)$ be such that $y(\theta_{-i}) = f(\theta_i, \theta_{-i})$ and $y^{'}(\theta_{-i}) = f(\theta_i^{'}, \theta_{-i})$, for all $\theta_{-i} \in \Theta_{-i}$. Clearly, $\sum_{\theta_{-i} \in \Theta_{-i}} z_i^1(\theta_{-i}) u_i\big(y(\theta_{-i}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i} \in \Theta_{-i}} z_i^1(\theta_{-i}) u_i\big(y^{'}(\theta_{-i}), (\theta_i, \theta_{-i})\big)$. □

Weak no-complete-indifference rules out indifference across payoff-type-*dependent* lotteries regardless of a player's belief about the payoff types of the other players. In contrast, it is easy to see that no-complete-indifference rules out indifference across payoff-type-*independent* lotteries.[16] Thus, the gap between environments where our characterization applies and those where RoRat-implementation (or RoEq-implementation) is feasible is limited to environments where some payoff type of some individual is indifferent between all payoff-type-independent lotteries for some first-order belief but that indifference does not extend to all payoff-type-dependent lotteries.[17] Are there any interesting environments that fall within that gap? This remains an open question.

[16] That is, no-complete-indifference is equivalent to the condition that for each $i \in I$, $\theta_i \in \Theta_i$, and $z_i^1 \in Z_i^1$, there exist $\ell, \ell^{'} \in \Delta(A)$ such that $\sum_{\theta_{-i} \in \Theta_{-i}} z_i^1(\theta_{-i}) u_i\big(\ell, (\theta_i, \theta_{-i})\big) \neq \sum_{\theta_{-i} \in \Theta_{-i}} z_i^1(\theta_{-i}) u_i\big(\ell^{'}, (\theta_i, \theta_{-i})\big)$.

[17] We have assumed that all individuals are relevant for the SCF f. If only a nonempty subset of individuals $I^* \subseteq I$ are relevant for the SCF f and f is RoRat-implementable (or RoEq-implementable), then the

9.4 Nonempty Best Response Property

We say the mechanism $\Gamma = \big((M_i)_{i\in I}, g\big)$ satisfies the *nonempty best response property* if best responses exist for all beliefs, i.e., $\arg\max_{m_i\in M_i} \psi_i(\theta_{-i}, m_{-i}) u_i\big(g(m_i, m_{-i}), (\theta_i, \theta_{-i})\big) \neq \emptyset$, for all $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$, $\theta_i \in \Theta_i$ and $i \in I$.

When the planner is restricted to mechanisms that satisfy the nonempty best response property, all the implementation notions discussed in this paper coincide. (This is true regardless of whether the environment satisfies no-complete-indifference or not.) Formally,

Proposition 9.6. *Suppose the mechanism $\Gamma = \big((M_i)_{i\in I}, g\big)$ satisfies the nonempty best response property. Then, for any SCF f,*

$$\Gamma \textit{ BfRat-implements } f \Leftrightarrow \Gamma \textit{ RoRat-implements } f$$
$$\Leftrightarrow \Gamma \textit{ Rat-implements } f \Leftrightarrow \Gamma \textit{ RoEq-implements } f.$$

Proof. We have already pointed out that the mechanism Γ RoEq-implements the SCF $f \Rightarrow$ Γ Rat-implements $f \Rightarrow \Gamma$ RoRat-implements $f \Rightarrow \Gamma$ BfRat-implements f.

Suppose the mechanism Γ satisfies the nonempty best response property. We only need to show that Γ BfRat-implements $f \Rightarrow \Gamma$ RoEq-implements f. Consider any type space $\mathcal{T}$.

Let σ be an interim equilibrium of the game $(\Gamma, \mathcal{T})$. Recall that belief-free rationalizability characterizes interim equilibria on all type spaces (Battigalli and Siniscalchi, 2003). Therefore, if $\sigma(t)[m] > 0$, then $m \in \mathcal{S}^\infty\big(\hat{\theta}(t)\big)$. As Γ BfRat-implements f, we have $g(m) = f(\theta)$.

We are therefore left to argue that the game $(\Gamma, \mathcal{T})$ has an interim equilibrium. It is sufficient to prove that Γ has an *ex post equilibrium*, i.e., there exists a profile $(s_i^*)_{i\in I}$, where $s_i^* : \Theta_i \to M_i$, for all $i \in I$, such that

$$u_i\big(g(s^*(\theta)), \theta\big) \geq u_i\big(g(m_i, s_{-i}^*(\theta_{-i})), \theta\big), \forall m_i \in M_i, \theta \in \Theta, i \in I.$$

Since Γ BfRat-implements f, we have $\mathcal{S}_i^\infty(\theta_i) \neq \emptyset$, for all $\theta_i \in \Theta_i$ and $i \in I$. Then for all $i \in I$, define s_i^* by fixing $s_i^*(\theta_i)$ to be any message in $\mathcal{S}_i^\infty(\theta_i)$, for all $\theta_i \in \Theta_i$.

Fix $i \in I$ and $\theta \in \Theta$. By construction, $s^*(\theta) \in \mathcal{S}^\infty(\theta)$. Since Γ BfRat-implements f, we have $g(s^*(\theta)) = f(\theta)$. Now, consider the belief $\psi_i \in \Delta(\Theta_{-i} \times M_{-i})$ such that $\psi_i\big(\theta_{-i}, s_{-i}^*(\theta_{-i})\big) = 1$. As Γ satisfies the nonempty best response property, there exists

environment must satisfy the weak no-complete-indifference condition that applies to only $i \in I^*$. In that case, as already mentioned in Footnote 6, all of our results can be obtained in environments that satisfy a similar weakening of no-complete-indifference. Thus, the gap between the environments where our results apply and those where RoRat-implementation (or RoEq-implementation) is feasible is still limited to the gap between indifference over all payoff-type-independent lotteries vs all payoff-type-dependent lotteries but now only for the relevant individuals.

$m_i^{'} \in \arg\max_{m_i \in M_i} \psi_i(\theta_{-i}^{'}, m_{-i}) u_i\big(g(m_i, m_{-i}), (\theta_i, \theta_{-i}^{'})\big)$. But since $\psi_i(\theta_{-i}^{'}, m_{-i}) > 0 \Rightarrow \theta_{-i}^{'} = \theta_{-i}$ and $m_{-i} = s_{-i}^*(\theta_{-i}) \in \mathcal{S}_{-i}^{\infty}(\theta_{-i})$, it follows that $m_i^{'} \in \mathcal{S}_i^{\infty}(\theta_i)$. Then, $g(m_i^{'}, s_{-i}^*(\theta_{-i})) = f(\theta)$ because Γ BfRat-implements f. As $g(s^*(\theta)) = g(m_i^{'}, s_{-i}^*(\theta_{-i}))$, it follows that $s_i^*(\theta_i) \in \arg\max_{m_i \in M_i} \psi_i(\theta_{-i}^{'}, m_{-i}) u_i\big(g(m_i, m_{-i}), (\theta_i, \theta_{-i}^{'})\big)$. Due to the construction of ψ_i, this means that $u_i\big(g(s^*(\theta)), \theta\big) \geq u_i\big(g(m_i, s_{-i}^*(\theta_{-i})), \theta\big)$, for all $m_i \in M_i$. Hence, $(s_i^*)_{i \in I}$ is an ex post equilibrium of Γ. □

Finite mechanisms satisfy the nonempty best response property. Thus, all notions of robust implementation discussed in this paper coincide when restricting to finite mechanisms. BM (2011) show that an additional "robust measurability" condition is necessary for robust implementation using finite mechanisms. Robust measurability is generally not related to weak RM.[18] It is, therefore, an additional restriction on robust implementation when using only finite mechanisms.

In single crossing aggregator (SCA) environments, BM (2007) show that, for responsive SCFs, robust measurability is equivalent to both the contraction property and strict RM. BM (2011, Section 5) suggest that, in SCA environments, robust implementation can be attained using the direct mechanism if the SCF satisfies strict EPIC and the contraction property. However, that claim is false in discrete settings, as demonstrated in Example 8.1. While it is in general difficult to obtain sufficient conditions for robust implementation using finite mechanisms, BM (2009b) show that EPIC and robust measurability are sufficient for robust *virtual* implementation using finite mechanisms.

In a complete-information environment with lotteries and transfers, Chen et al. (2021) 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 et al., 2022). 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.

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

9.5 Ex Post Best Response Property

The mechanism $\Gamma = \big((M_i)_{i\in I}, g\big)$ satisfies the *ex post best response property* if for all $i \in I$ and $\theta_i \in \Theta_i$, there exists a message $m_i^*(\theta_i) \in \mathcal{S}_i^\infty(\theta_i)$ such that

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

for all $\theta_{-i} \in \Theta_{-i}$ and $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$. This notion was introduced by Bergemann and Morris (2011).

Notice that the ex post best response property is a condition imposed on the belief-free rationalizability correspondence, $\mathcal{S}^\infty$. Thus, unlike the nonempty best response property, which is guaranteed to hold under natural restrictions (e.g., finite mechanisms or compact mechanisms when utility functions are continuous), we need to first determine the belief-free rationalizability correspondence of a mechanism in order to check whether the mechanism satisfies the ex post best response property. To the extent that the literature has given a reasonable amount of attention to the ex post best response property, we find it useful to discuss its implications in robust implementation below.

In general, the ex post best response property is not related to the nonempty best response property. But, in the context of robust implementation, nonempty best response property implies the ex post best response property in the following sense:

Lemma 9.7. *Suppose the mechanism $\Gamma = \big((M_i)_{i\in I}, g\big)$ BfRat-implements the SCF f.*[19] *If Γ satisfies the nonempty best response property, then it satisfies the ex post best response property.*

Proof. Since Γ BfRat-implements f and satisfies the nonempty best response property, Γ has an ex post equilibrium $(s_i^*)_{i\in I}$ such that $s_i^*(\theta_i) \in \mathcal{S}_i^\infty(\theta_i)$, for all $\theta_i \in \Theta_i$ and $i \in I$ (as argued in the proof of Proposition 9.6).

Fix $i \in I$ and $\theta_i \in \Theta_i$. Since Γ satisfies the nonempty best response property, we have $\arg\max_{m_i\in M_i} u_i\big(g(m_i, m_{-i}), (\theta_i, \theta_{-i})\big) \neq \emptyset$, for all $\theta_{-i} \in \Theta_{-i}$ and $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$.

For any $\theta_{-i} \in \Theta_{-i}$ and $m_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})$, pick $m_i^{'} \in \arg\max_{m_i\in M_i} u_i\big(g(m_i, m_{-i}), (\theta_i, \theta_{-i})\big)$. Then, $(m_i^{'}, m_{-i}) \in \mathcal{S}^\infty(\theta)$. But we also have $(s_i^*(\theta_i), m_{-i}) \in \mathcal{S}^\infty(\theta)$. Then it follows that $g(s_i^*(\theta_i), m_{-i}) = g(m_i^{'}, m_{-i}) = f(\theta)$ as Γ BfRat-implements f. So, we have $s_i^*(\theta_i) \in \arg\max_{m_i\in M_i} u_i\big(g(m_i, m_{-i}), (\theta_i, \theta_{-i})\big)$.

The ex post property is then satisfied by setting $m_i^*(\theta_i) = s_i^*(\theta_i)$, for all $\theta_i \in \Theta_i$ and $i \in I$. □

[19] Since BfRat-implementation is the weakest notion of implementation discussed in this paper, the lemma holds for all the other implementation notions too.

The converse of the above lemma is not true. For example, the canonical mechanism in BM (2011) RoEq-implements (and hence, BfRat-implements) an SCF and satisfies the ex post best response property but fails the nonempty best response property because it includes an integer game.

Recall from BM (2011, Theorem 3) that if a mechanism RoEq-implements an SCF, then the *same* mechanism Rat-implements the SCF; and the converse is true if the mechanism satisfies the ex post best response property. Thus, RoEq-implementation and Ra-implementation coincide when the planner is restricted to mechanisms that satisfy the ex post best response property. We now show that, in fact, all the implementation notions discussed in this paper coincide under that restriction. (This is true regardless of whether the environment satisfies no-complete-indifference or not.) Formally,

Proposition 9.8. *Suppose the mechanism* $\Gamma = \big((M_i)_{i\in I}, g\big)$ *satisfies the ex post best response property. Then, for any SCF* f*,*

$$\Gamma \textit{ BfRat-implements } f \Leftrightarrow \Gamma \textit{ RoRat-implements } f$$
$$\Leftrightarrow \Gamma \textit{ Rat-implements } f \Leftrightarrow \Gamma \textit{ RoEq-implements } f.$$

Proof. As in the proof of Proposition 9.6, we only need to show that if the mechanism Γ satisfies the ex post best response property and BfRat-implements the SCF f, then Γ RoEq-implements f. Like in that proof, the uniqueness requirement of RoEq-implementation is satisfied because Γ BfRat-implements f and belief-free rationalizability characterizes interim equilibria on all type spaces. Next, since Γ satisfies the ex post best response property, Γ has an ex post equilibrium.[20] Therefore, Γ has a pure-strategy interim equilibrium in all type spaces, which completes the argument.[21] □

10 Conclusion

We examine the implications of (interim correlated) rationalizability as the solution concept in robust implementation. The resulting notion, RoRat-implementation, is connected but *not* equivalent to implementation in belief-free rationalizability (BfRat-implementaion). Indeed,

[20] If Γ satisfies the ex post best response property, then it is straightforward to see that the profile $(s_i^*)_{i\in I}$, where $s_i^*(\theta_i) = m_i^*(\theta_i)$, for all $\theta_i \in \Theta_i$, forms an ex post equilibrium of Γ.

[21] This qualifies the main result in Jain et al. (2023). They argue that for a fixed mechanism Γ, wr-implementation, s-implementation, and Rat-implementation are strictly nested. That is, Γ wr-implements $f \Rightarrow \Gamma$ s-implements $f \Rightarrow \Gamma$ Rat-implements f but the converse of neither implication is true. But, as we have argued, if Γ satisfies the ex post best response property, all these implementation notions coincide: Γ wr-implements $f \Leftrightarrow \Gamma$ s-implements $f \Leftrightarrow \Gamma$ Rat-implements f.

we show that RoRat-implementation is equivalent to wRat-implementation, which is stronger than BfRat-implementation because it requires that best responses to all first-order beliefs exist. Utilizing this equivalence, we prove that weak RM characterizes RoRat-implementation in environments satisfying the mild no-complete-indifference condition. Weak RM can be decomposed into incentive and monotonicity-type constraints: precisely, weak RM is equivalent to semi-strict EPIC and the preference-reversal condition. Interestingly, semi-strict EPIC and the preference-reversal condition coincide in private-value environments but not more generally.

We also clarify the relationships between different "robust" and "rationalizable" implementation notions discussed in the literature. We prove that strict RM characterizes both RoEq-implementation and Rat-implementation under our mild restriction on the environments, closing the gap not only between the necessary and sufficient conditions for RoEq-implementation but also between RoEq-implementation and Rat-implementation in BM (2011). Strict RM can be strictly stronger than weak RM; thus, RoEq-implementation can be strictly stronger than RoRat-implementation. Hence, equilibrium and rationalizablity have different implications for robust implementation when the designer is unrestricted in the choice of mechanisms. However, when restricted to mechanisms satisfying the nonempty best response property, the set of robustly implementable SCFs are the same regardless of whether we adopt equilibrium, rationalizability, or belief-free rationalizability as the solution concept. The characterization of these SCFs remains an open question for future research.

11 Appendix

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

11.1 BfRat-Implementation and wRat-Implementation

Recall that RoRat-implementation is equivalent to wRat-implementation (Theorem 3.4). Although the uniqueness requirement of wRat-implementation is identical to that of BfRat-implementation, the nonemptiness requirement for wRat-implementation is significantly more involved than that for BfRat-implementation. In what follows, we argue by means of an example that the nonemptiness requirement of wRat-implementation results in a substantial constraint on implementability whereas BfRat-implementation is still permissive.

Suppose that there are two agents, $I = \{1, 2\}$; three alternatives, $A = \{a, b, c\}$; agent 1's payoff type space, $\Theta_1 = \{\theta_1, \theta_1'\}$; and agent 2's payoff type space, $\Theta_2 = \{\theta_2, \theta_2'\}$. The SCF f

is such that:

f	θ_2	θ_2'
θ_1	a	$\frac{1}{2}b + \frac{1}{2}c$
θ_1'	c	b

The environment is one of *private values* with utilities as follows:

(θ_1, θ_2)	a	b	c
Agent 1	4	2	3
Agent 2	2	3	4

(θ_1, θ_2')	a	b	c
Agent 1	4	2	3
Agent 2	0	4	2

(θ_1', θ_2)	a	b	c
Agent 1	2	3	4
Agent 2	2	3	4

(θ_1', θ_2')	a	b	c
Agent 1	2	3	4
Agent 2	0	4	2

We claim that the SCF f is *not* wRat-implementable. The SCF f does not satisfy EPIC, which is implied by weak RM, the key necessary condition for RoRat-implementation (See Lemma 5.2 for this). To see the violation of EPIC, since player 1 of payoff type θ_1' finds c better than b, we have

$$u_1(f(\theta_1', \theta_2'), \theta_1') < u_1(f(\theta_1, \theta_2'), \theta_1').$$

Thus, f is *not* RoRat-implementable. By the equivalence between RoRat-implementation and wRat-implementation, we conclude that f is *not* wRat-implementable.

Using the same example, however, we show that the SCF f is BfRat-implementable. Consider the mechanism with $M_1 = \{m_1^{-1}, m_1^0, m_1^1, m_1^2, \dots\}$, $M_2 = \{m_2^{-1}, m_2^0, m_2^1, m_2^2, \dots\}$, and the outcome function g as follows:

$g(m)$	m_2^{-1}	m_2^0	m_2^1	m_2^2	m_2^3	$\cdots$
m_1^{-1}	b	c	$\frac{1}{3}b + \frac{2}{3}c$	$\frac{1}{3}b + \frac{2}{3}c$	$\frac{1}{3}b + \frac{2}{3}c$	$\cdots$
m_1^0	$\frac{1}{2}b + \frac{1}{2}c$	a	$\frac{1}{2}a + \frac{1}{2}c$	$\frac{1}{3}a + \frac{2}{3}c$	$\frac{1}{4}a + \frac{3}{4}c$	$\cdots$
m_1^1	$\frac{1}{2}b + \frac{1}{2}c$	$\frac{1}{4}b + \frac{3}{4}c$	$\frac{1}{3}b + \frac{2}{3}c$	$\frac{1}{3}b + \frac{2}{3}c$	$\frac{1}{3}b + \frac{2}{3}c$	$\cdots$
m_1^2	$\frac{1}{3}b + \frac{2}{3}c$	$\frac{1}{5}b + \frac{4}{5}c$	$\frac{1}{4}b + \frac{3}{4}c$	$\frac{1}{4}b + \frac{3}{4}c$	$\frac{1}{4}b + \frac{3}{4}c$	$\cdots$
m_1^3	$\frac{1}{4}b + \frac{3}{4}c$	$\frac{1}{6}b + \frac{5}{6}c$	$\frac{1}{5}b + \frac{4}{5}c$	$\frac{1}{5}b + \frac{4}{5}c$	$\frac{1}{5}b + \frac{4}{5}c$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

First, consider the first step of eliminating never best responses according to b^Θ, the best response operator for payoff types.

1. We argue that $\mathcal{S}_1^1(\theta_1) = \{m_1^0\}$. Pick m_1^z such that $z \geq 1$. The message m_1^z is strictly dominated by message m_1^{z+1} because c is better than b for payoff type θ_1. Hence, m_1^z is never a best response to any belief. Now consider message m_1^{-1}. Since a is better than c, which in turn is better than b for payoff type θ_1, message m_1^{-1} is strictly dominated by message m_1^0. Hence, m_1^{-1} is never a best response to any belief. Finally, since a is the most-preferred outcome for payoff type θ_1, message m_1^0 is a best response to any belief $\psi_1 \in \Delta(\Theta_2 \times M_2)$ such that player 2 plays m_2^0 with probability 1. Thus, $\mathcal{S}_1^1(\theta_1) = \{m_1^0\}$.

2. We argue that $\mathcal{S}_1^1(\theta_1^{'}) = \{m_1^{-1}\}$. Pick m_1^z such that $z \geq 1$. The message m_1^z is strictly dominated by message m_1^{z+1} because c is better than b for payoff type $\theta_1^{'}$. Hence, m_1^z is never a best response to any belief.

   Next, consider message m_1^0 and pick any belief $\psi_1 \in \Delta(\Theta_2 \times M_2)$. Let $\hat{\psi}_1$ denote the marginal of ψ_1 on M_2. Then the expected payoff of payoff type $\theta_1^{'}$ when she plays m_1^0 and holds the belief ψ_1 is

$$\hat{\psi}_1(m_2^{-1})(3.5) + \hat{\psi}_1(m_2^0)(2) + \sum_{z^{'} \geq 1} \hat{\psi}_1(m_2^{z^{'}}) \left(\frac{1}{z^{'}+1}(2) + \frac{z^{'}}{z^{'}+1}(4) \right) < 4.$$

   If instead she were to play m_1^z, then her expected payoff is

$$\hat{\psi}_1(m_2^{-1}) \left(\frac{1}{z+1}(3) + \frac{z}{z+1}(4) \right) + \hat{\psi}_1(m_2^0) \left(\frac{1}{z+3}(3) + \frac{z+2}{z+3}(4) \right)$$
$$\left(\frac{1}{z+2}(3) + \frac{z+1}{z+2}(4) \right) \sum_{z^{'} \geq 1} \hat{\psi}_1(m_2^{z^{'}}).$$

   Since the limit of the above expression as $z \to \infty$ is equal to 4, we conclude that there is a sufficiently high enough z such that m_1^z is a better response than m_1^0 when player 1 of payoff type $\theta_1^{'}$ holds the belief ψ_1. Thus, m_1^0 is never a best response to any belief.

   Finally, since c is the most-preferred outcome for payoff type $\theta_1^{'}$, message m_1^{-1} is a best response to any belief $\psi_1 \in \Delta(\Theta_2 \times M_2)$ such that player 2 plays m_2^0 with probability 1. Thus, $\mathcal{S}_1^1(\theta_1^{'}) = \{m_1^{-1}\}$.

3. We argue that $\mathcal{S}_2^1(\theta_2) = \{m_2^0\}$. Pick m_2^z such that $z \geq 1$ and any belief $\psi_2 \in \Delta(\Theta_1 \times M_1)$. Let $\hat{\psi}_2$ denote the marginal of ψ_2 on M_1. First, suppose $\hat{\psi}_2(m_1^0) = 0$. Since c is better than b for payoff type θ_2, message m_2^0 gives a higher expected payoff than message m_2^z. Second, suppose $\hat{\psi}_2(m_1^0) > 0$. Then, since c is better than a for payoff type θ_2, message m_2^{z+1} gives a higher expected payoff than message m_2^z. Hence, m_2^z is

never a best response to any belief.

Next, consider message m_2^{-1}. For payoff type θ_2, lottery $1/5a + 4/5c$ is better than $1/2b+1/2c$ and alternative c is better than b. Hence, message m_2^{-1} is strictly dominated by message m_2^4. So, m_2^{-1} is never a best response to any belief.

Finally, since c is the most-preferred alternative for payoff type θ_2, message m_2^0 is a best response to any belief $\psi_2 \in \Delta(\Theta_1 \times M_1)$ such that player 1 plays m_1^{-1} with probability 1. Thus, $\mathcal{S}_2^1(\theta_2) = \{m_2^0\}$.

4. We argue that $\mathcal{S}_2^1(\theta_2') = \{m_2^{-1}\}$. This is because every message $m_2 \neq m_2^{-1}$ is strictly dominated by m_2^{-1} for payoff type θ_2' since b is better than c, which in turn is better than a for payoff type θ_2'.

   Since b is the most-preferred alternative for payoff type θ_2', m_2^{-1} is a best response to any belief $\psi_2 \in \Delta(\Theta_1 \times M_1)$ such that player 1 plays m_1^{-1} with probability 1. Thus, $\mathcal{S}_2^1(\theta_2') = \{m_2^{-1}\}$.

Next, consider the second step of eliminating never best responses according to b^Θ.

1. $\mathcal{S}_1^2(\theta_1) = \{m_1^0\}$. This is because if player 1's belief is such that $\psi_1(\theta_2, m_2^0) = 1$, then indeed it is a best response for player 1 of payoff type θ_1 to play m_1^0.

2. $\mathcal{S}_1^2(\theta_1') = \{m_1^{-1}\}$. This is because if player 1's belief is such that $\psi_1(\theta_2, m_2^0) = 1$, then indeed it is a best response for player 1 of payoff type θ_1' to play m_1^{-1}.

3. $\mathcal{S}_2^2(\theta_2) = \{m_2^0\}$. This is because if player 2's belief is such that $\psi_2(\theta_1', m_1^{-1}) = 1$, then indeed it is a best response for player 2 of payoff type θ_2 to play m_2^0.

4. $\mathcal{S}_2^2(\theta_2') = \{m_2^{-1}\}$. This is because if player 2's belief is such that $\psi_2(\theta_1', m_1^{-1}) = 1$, then indeed it is a best response for player 2 of payoff type θ_2' to play m_2^{-1}.

Since the first two steps of elimination coincide, we have $\mathcal{S}_1^\infty(\theta_1) = \{m_1^0\}$, $\mathcal{S}_1^\infty(\theta_1') = \{m_1^{-1}\}$, $\mathcal{S}_2^\infty(\theta_2) = \{m_2^0\}$, and $\mathcal{S}_2^\infty(\theta_2') = \{m_2^{-1}\}$. Furthermore, if $m \in \mathcal{S}^\infty(\hat{\theta})$, then the outcome is $f(\hat{\theta})$, thereby completing the argument.

11.2 Omitted Proofs

Proof of Theorem 4.3.

($\Rightarrow$): We first show that weak RM is necessary for RoRat-implementation.

Suppose the mechanism $\Gamma = ((M_i)_{i\in I}, g)$ RoRat-implements f. It follows from Theorem 3.4 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 $\tilde{m}_i \in M_i$,

$$\begin{aligned}
&\sum_{\tilde{\theta}_{-i}, \tilde{m}_{-i}} \psi_i^\theta(\tilde{\theta}_{-i}, \tilde{m}_{-i}) u_i\big(g(\tilde{m}_i, \tilde{m}_{-i}), (\theta_i, \tilde{\theta}_{-i})\big) \\
&= \sum_{\tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})} \text{marg}_{M_{-i}} \psi_i^\theta(\tilde{m}_{-i}) u_i\big(g(\tilde{m}_i, \tilde{m}_{-i}), \theta\big) \\
&= u_i\left(\sum_{\tilde{m}_{-i} \in \mathcal{S}_{-i}^\infty(\theta_{-i})} \text{marg}_{M_{-i}} \psi_i^\theta(\tilde{m}_{-i}) g(\tilde{m}_i, \tilde{m}_{-i}), \theta \right). \qquad (4)
\end{aligned}$$

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

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

where the first and last equality follows from (4). 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 $\sum_{\tilde{m}_{-i}\in\mathcal{S}^{\infty}_{-i}(\theta_{-i})} \text{marg}_{M_{-i}}\psi^{\theta}_i(\tilde{m}_{-i})g(\tilde{m}_i,\tilde{m}_{-i}) = \ell$. Then, as per the above arguments, $u_i(\ell,\theta) \geq u_i(f(\theta),\theta)$ is equivalent to

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

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

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

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 $\theta'_i \not\sim^f_i \theta_i$, 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

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i})u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big) \geq \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i})u_i\big(y(\theta'_{-i}),(\theta_i,\theta_{-i})\big). \qquad (5)$$

We first show that for any $i \in I$, $\theta_i \in \Theta_i$, and $\theta'_i \in \beta_i(\theta_i)$ satisfying $\theta'_i \sim^f_i \theta_i$, 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 (5) holds for all $y \in Y_i[\tilde{\theta}_i]$.

Pick any i, θ_i, and $\theta'_i \in \beta_i(\theta_i)$ satisfying $\theta'_i \sim^f_i \theta_i$. 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 $\theta_i \sim^f_i \theta'_i$, 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 (5) 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 (5) 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, (5) holds for all $y \in Y_i[\tilde{\theta}_i]$. Hence, for any $m_i \in M_i$,

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big) \geq \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y^{m_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big). \qquad (6)$$

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

$$\begin{aligned}
& \sum_{\theta_{-i},m_{-i}} \psi_i^\Gamma(\theta_{-i},m_{-i}) u_i\big(g(m'_i,m_{-i}),(\theta_i,\theta_{-i})\big) \\
= & \sum_{\theta_{-i},m_{-i}} \left(\sum_{\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) \times \mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}) u_i\big(g(m'_i,m_{-i}),(\theta_i,\theta_{-i})\big) \right) \\
& \quad \text{(by definition of } \psi_i^\Gamma) \\
= & \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big) \\
& \quad \left(\begin{array}{l} \text{by weak rationalizable implementability of } f \text{ because } m'_i \in \mathcal{S}_i^\infty(\theta'_i) \\ \text{and } \mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}) > 0 \text{ implies } m_{-i} \in \mathcal{S}_{-i}^\infty(\theta'_{-i}) \end{array} \right) \\
\geq & \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) u_i\big(y^{m_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big) \\
& \quad (\because \text{inequality (6) holds for any } m_i \in M_i) \\
= & \sum_{\theta_{-i},\theta'_{-i}} \psi_i(\theta_{-i},\theta'_{-i}) \left(\sum_{m_{-i}} \mathrm{marg}_{M_{-i}} \psi_i^{(\tilde{\theta}_i,\theta'_{-i})}(m_{-i}) u_i\big(g(m_i,m_{-i}),(\theta_i,\theta_{-i})\big) \right) \\
& \quad \text{(by definition of } y^{m_i}) \\
= & \sum_{\theta_{-i},m_{-i}} \psi_i^\Gamma(\theta_{-i},m_{-i}) u_i\big(g(m_i,m_{-i}),(\theta_i,\theta_{-i})\big) \\
& \quad \text{(by definition of } \psi_i^\Gamma).
\end{aligned}$$

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 of necessity. □

($\Leftarrow$): We use the mechanism Γ constructed in Section 4 to prove that Γ wRat-implements f, which implies that Γ RoRat-implements f because of Theorem 3.4. We first note two useful technical lemmata.

Lemma 11.1. *If the SCF f satisfies weak RM, then for all $i \in I$, $\theta_i, \tilde{\theta}_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$, there exists $y \in Y_i^*[\tilde{\theta}_i]$ such that*

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

Proof. Suppose the SCF f satisfies weak RM. By Lemma 5.2, f satisfies semi-strict EPIC. Pick any $i \in I$, $\theta_i, \tilde{\theta}_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$. By applying the same argument as in the proof of Lemma 7.10, we can find $y^\epsilon : \Theta_{-i} \to \Delta(A)$ and $y'^\epsilon : \Theta_{-i} \to \Delta(A)$ such that

$$u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > u_i\big(y^\epsilon(\theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) \text{ and } u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > u_i\big(y'^\epsilon(\theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big),$$

for all $\theta'_{-i} \in \Theta_{-i}$, and

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

Since $\Delta^*(A)$ is a dense subset of $\Delta(A)$, for each θ'_{-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 θ'_{-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 θ'_{-i}. As Θ_{-i} is finite, there exists a sufficiently large z such that

$$u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > u_i\big(y^z(\theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) \text{ and } u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > u_i\big(y'^z(\theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big),$$

for all θ'_{-i}, and

$$\sum_{\theta_{-i}, \theta'_{-i}} \psi_i(\theta_{-i}, \theta'_{-i}) u_i\big(y^z(\theta'_{-i}), (\theta_i, \theta_{-i})\big) > \sum_{\theta_{-i}, \theta'_{-i}} \psi_i(\theta_{-i}, \theta'_{-i}) u_i\big(y'^z(\theta'_{-i}), (\theta_i, \theta_{-i})\big). \tag{7}$$

The first set of inequalities imply that $y^z, y'^z \in Y_i^*[\tilde{\theta}_i]$.

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

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

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 (7). □

Lemma 11.2. *For all $i \in I$, $\theta_i \in \Theta_i$ and $z_i^1 \in Z_i^1$, there exists $a \in A$ such that*

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

Proof. 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 $a, a^{'} \in A$, which contradicts no-complete-indifference.. □

Now, we are ready to the prove that Γ wRat-implements f. The proof 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

$$\left(\frac{m_i^2}{1+m_i^2}\right) \sum_{\theta_{-i}, \theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) u_i\big(m_i^3[\tilde{\theta}_i](\theta'_{-i}), (\theta_i, \theta_{-i})\big)$$
$$+ \left(1 - \frac{m_i^2}{1+m_i^2}\right) \sum_{\theta_{-i}, \theta'_{-i}} \psi_i^{2,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) u_i\big(y_i^{\tilde{\theta}_i}(\theta'_{-i}), (\theta_i, \theta_{-i})\big). \tag{8}$$

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

$$\frac{1}{n}\left(\frac{m_i^2}{1+m_i^2}\right) \sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(m_i^4, (\theta_i, \theta_{-i})\big) + \frac{1}{n}\left(1 - \frac{m_i^2}{1+m_i^2}\right) \sum_{\theta_{-i}} \psi_i^3(\theta_{-i}) u_i\big(\bar{\alpha}, (\theta_i, \theta_{-i})\big)$$
$$+ \sum_{\theta_{-i}, m_{-i} \in M^3_{-i}} \frac{\psi_i(\theta_{-i}, m_{-i})}{\Psi_i^3} \sum_{j \neq i} \left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right) u_i\big(m_j^4, (\theta_i, \theta_{-i})\big) + \frac{1}{n}\left(1 - \frac{m_j^2}{1+m_j^2}\right) u_i\big(\bar{\alpha}, (\theta_i, \theta_{-i})\big)\right). \tag{9}$$

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 11.1.

▷ 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 11.2.

  ▷ 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 (8) 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 (9) 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 $\theta_i^{'}\not\sim_i^f\theta_i$ 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

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

As $\theta_i^{'}\not\sim_i^f\theta_i$ 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

$$\sum_{\theta_{-i},\theta'_{-i}} \psi_i^{1,\tilde{\theta}_i}(\theta_{-i}, \theta'_{-i}) u_i\big(f(\theta'_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big). \tag{10}$$

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 (10) 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 3.4 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 sufficiency. □

Proof of Proposition 5.7.

($\Leftarrow$) Pick an unacceptable deception β. Then, according to the preference-reversal condition, 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 $\tilde{\theta}_i \sim_i^f \theta_i^{'}$, there exists $y_{\tilde{\theta}_i} \in Y_i^w[\tilde{\theta}_i]$ such that

$$u_i\big(y_{\tilde{\theta}_i}(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) > u_i\big(f(\tilde{\theta}_i, \theta_{-i}^{'}), (\theta_i, \theta_{-i})\big),$$

for all $\theta_{-i} \in \Theta_{-i}$ and $\theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$.

Pick any $\tilde{\theta}_i \in \Theta_i$ and any belief $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ satisfying $\psi_i(\theta_{-i}, \theta_{-i}^{'}) > 0 \Rightarrow \theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$.

We divide the rest of the argument into the following two cases.

Case 1: $\tilde{\theta}_i \not\sim_i^f \theta_i^{'}$.

Trivially, $\theta_i^{'} \sim_i^f \theta_i^{'}$. Thus, by the preference-reversal condition, there exists $y_{\theta_i^{'}} \in Y_i^w[\theta_i^{'}]$ such that for any $\theta_{-i} \in \Theta_{-i}$ and $\theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$,

$$u_i\big(y_{\theta_i^{'}}(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) > u_i\big(f(\theta_i^{'}, \theta_{-i}^{'}), (\theta_i, \theta_{-i})\big). \tag{11}$$

Pick any $\varepsilon \in (0,1)$. For each $\theta_{-i}^{'} \in \Theta_{-i}$, we define

$$y^{\varepsilon}(\theta_{-i}^{'}) \equiv \varepsilon y_{\theta_i^{'}}(\theta_{-i}^{'}) + (1-\varepsilon) f(\theta_i^{'}, \theta_{-i}^{'}).$$

Since $\tilde{\theta}_i \not\sim_i^f \theta_i^{'}$, by semi-strict EPIC, we have that, for any $\theta_{-i}^{'} \in \Theta_{-i}$,

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

Since Θ_{-i} is finite, we can find $\varepsilon \in (0,1)$ small enough so that, for any $\theta_{-i}^{'} \in \Theta_{-i}$,

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

This further implies that $y^{\varepsilon} \in Y_i[\tilde{\theta}_i]$. In the rest of the argument, we fix any such small enough $\varepsilon \in (0,1)$.

Since $\varepsilon \in (0,1)$, it follows from (11) and the definition of y^{ε} that, for any $\theta_{-i} \in \Theta_{-i}$ and $\theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$,

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

This further implies that

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

Case 2: $\tilde{\theta}_i \sim_i^f \theta'_i$.

Since $\tilde{\theta}_i \sim_i^f \theta'_i$, by the preference-reversal condition, there exists $y_{\tilde{\theta}_i} \in Y_i^w[\tilde{\theta}_i]$ such that, for any $\theta_{-i} \in \Theta_{-i}$ and $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$,

$$\begin{aligned} u_i\big(y_{\tilde{\theta}_i}(\theta'_{-i}),(\theta_i,\theta_{-i})\big) &> u_i\big(f(\tilde{\theta}_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big) \\ &= u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big) \ (\because\ \tilde{\theta}_i \sim_i^f \theta'_i). \end{aligned} \tag{12}$$

Pick any $\varepsilon \in (0,1)$. For each $\theta'_{-i} \in \Theta_{-i}$, we define

$$y^\varepsilon(\theta'_{-i}) \equiv \varepsilon y_{\tilde{\theta}_i}(\theta'_{-i}) + (1-\varepsilon) f(\theta_i,\theta'_{-i}).$$

Since $\theta'_i \not\sim_i^f \theta_i$ but $\tilde{\theta}_i \sim_i^f \theta'_i$, it follows from the transitivity of $\sim_i^f$ that $\tilde{\theta}_i \not\sim_i^f \theta_i$. Then, by semi-strict EPIC, we have that, for any $\theta'_{-i} \in \Theta_{-i}$,

$$u_i\big(f(\tilde{\theta}_i,\theta'_{-i}),(\tilde{\theta}_i,\theta'_{-i})\big) > u_i\big(f(\theta_i,\theta'_{-i}),(\tilde{\theta}_i,\theta'_{-i})\big).$$

Since $y_{\tilde{\theta}_i} \in Y_i^w[\tilde{\theta}_i]$ and $\varepsilon \in (0,1)$, it follows from the definition of y^ε that, for any $\theta'_{-i} \in \Theta_{-i}$,

$$u_i\big(f(\tilde{\theta}_i,\theta'_{-i}),(\tilde{\theta}_i,\theta'_{-i})\big) > u_i\big(y^\varepsilon(\theta'_{-i}),(\tilde{\theta}_i,\theta'_{-i})\big).$$

This implies that $y^\varepsilon \in Y_i[\tilde{\theta}_i]$.

Since $\Theta_{-i} \times \Theta_{-i}$ is finite, it follows from (12) and the definition of y^ε that we can find $\varepsilon \in (0,1)$ large enough so that, for any $\theta_{-i} \in \Theta_{-i}$ and $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$,

$$u_i\big(y^\varepsilon(\theta'_{-i}),(\theta_i,\theta_{-i})\big) > u_i\big(f(\theta'_i,\theta'_{-i}),(\theta_i,\theta_{-i})\big).$$

Fixing any such large enough $\varepsilon \in (0,1)$, the above implies that

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

After considering the two cases, we thus conclude that the SCF f satisfies weak RM. □

($\Rightarrow$) Suppose the SCF f satisfies weak RM. Lemma 5.2 shows that f satisfies semi-strict

EPIC. We now argue that f satisfies the preference-reversal condition.

Pick any unacceptable deception β. Then, by weak RM, 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

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

Fix $\tilde{\theta}_i \sim_i^f \theta_i^{'}$ and $\theta_{-i}^{'} \in \Theta_{-i}$. The set of $\theta_{-i} \in \Theta_{-i}$ such that $\theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$ is nonempty because $\theta_{-i}^{'} \in \beta_{-i}(\theta_{-i}^{'})$. Pick any first-order belief $z_i^1 \in Z_i^1$ satisfying $z_i^1(\theta_{-i}) > 0 \Rightarrow \theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})$. Consider the belief $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ such that $\psi_i(\theta_{-i}, \tilde{\theta}_{-i}) = z_i^1(\theta_{-i})$, if $\tilde{\theta}_{-i} = \theta_{-i}^{'}$, and $\psi_i(\theta_{-i}, \tilde{\theta}_{-i}) = 0$, otherwise. By weak RM, there exists $y^{'}(\theta_{-i}^{'}) \in \Delta(A)$ such that

$$u_i\big(f(\tilde{\theta}_i, \theta_{-i}^{'}), (\tilde{\theta}_i, \theta_{-i}^{'})\big) > u_i\big(y^{'}(\theta_{-i}^{'}), (\tilde{\theta}_i, \theta_{-i}^{'})\big) \tag{13}$$

and

$$\begin{aligned}\sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(y^{'}(\theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) &> \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(f(\theta_i^{'}, \theta_{-i}^{'}), (\theta_i, \theta_{-i})\big) \\ &= \sum_{\theta_{-i}} z_i^1(\theta_{-i}) u_i\big(f(\tilde{\theta}_i, \theta_{-i}^{'}), (\theta_i, \theta_{-i})\big), \end{aligned} \tag{14}$$

where the equality follows from the fact that $\tilde{\theta}_i \sim_i^f \theta_i^{'}$.

Consider the normal-form game with two players i and k such that the set of actions for players i and k are equal to

$$\hat{M}_i = \big\{\ell \in \Delta(\bar{A}) : u_i\big(f(\tilde{\theta}_i, \theta_{-i}^{'}), (\tilde{\theta}_i, \theta_{-i}^{'})\big) \geq u_i\big(\ell, (\tilde{\theta}_i, \theta_{-i}^{'})\big)\big\},$$

and

$$\hat{M}_k = \{\theta_{-i} \in \Theta_{-i} : \theta_{-i}^{'} \in \beta_{-i}(\theta_{-i})\},$$

respectively. The payoff function for player i is given by $\hat{u}_i(\ell, \theta_{-i}) = u_i\big(\ell, (\theta_i, \theta_{-i})\big)$. Finally, the payoff function for player k is given by $\hat{u}_k(\ell, \theta_{-i}) = -\hat{u}_i(\ell, \theta_{-i})$. This makes it a two-person zero-sum game.

$\Delta(\bar{A})$ is compact in the weak* topology because $\bar{A}$ is compact. Hence, $\hat{M}_i$ is a closed subset of $\Delta(\bar{A})$, and therefore, compact in the weak* topology. As Θ_{-i} is finite, $\hat{M}_k$ is compact in the discrete topology. In addition, the payoff functions are continuous.

It follows from (13) and (14) that the action $f(\tilde{\theta}_i, \theta_{-i}^{'})$ is never a best response for player i in the normal-form game. By applying the same arguments as in Pearce (1984, Lemma 3)

and using the fact that the action spaces are compact and payoff functions are continuous – which guarantees the existence of Nash equilibrium –, we can establish that the action $f(\tilde{\theta}_i, \theta'_{-i})$ is strictly dominated for player i in the normal-form game. Hence, there exists a strategy $\sigma_i \in \Delta(\hat{M}_i)$ of player i in the normal-form game such that

$$\int_{\hat{M}_i} \hat{u}_i(\hat{m}_i, \theta_{-i})\sigma_i(d\hat{m}_i) > \hat{u}_i\big(f(\tilde{\theta}_i, \theta'_{-i}), \theta_{-i}\big),$$

for all $\theta_{-i} \in \hat{M}_k$. Observe that $\hat{M}_i$ is clearly convex. By the linearity of the expected utility function and convexity of $\hat{M}_i$, there exists $\bar{m}_i \in \hat{M}_i$ such that

$$\hat{u}_i(\bar{m}_i, \theta_{-i}) = \int_{\hat{M}_i} \hat{u}_i(\hat{m}_i, \theta_{-i})\sigma_i(d\hat{m}_i) > \hat{u}_i\big(f(\tilde{\theta}_i, \theta'_{-i}), \theta_{-i}\big), \tag{15}$$

for all $\theta_{-i} \in \hat{M}_k$. As $\bar{m}_i \in \hat{M}_i$, we also have $u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) \geq u_i\big(\bar{m}_i, (\tilde{\theta}_i, \theta'_{-i})\big)$.

Recall Equation (13) involving $y'(\theta'_{-i})$. Pick any $\varepsilon \in (0, 1)$, and let $\bar{m}_i^{\varepsilon} = (1-\varepsilon)\bar{m}_i + \varepsilon y'(\theta'_{-i})$. Since Θ_{-i} is finite, it follows from (13) and (15) that there exists $\varepsilon > 0$ small enough such that

$$\hat{u}_i(\bar{m}_i^{\varepsilon}, \theta_{-i}) > \hat{u}_i\big(f(\tilde{\theta}_i, \theta'_{-i}), \theta_{-i}\big),$$

for all $\theta_{-i} \in \hat{M}_k$ and $u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > u_i\big(\bar{m}_i^{\varepsilon}, (\tilde{\theta}_i, \theta'_{-i})\big)$.

The set of probability measures with a finite support is dense in $\Delta(\bar{A})$. Since the utility function is continuous, Θ_{-i} is finite, and $\bar{A}$ is the closure of A, there exists $\bar{m}'_i \in \Delta(A)$ with finite support such that

$$\hat{u}_i(\bar{m}'_i, \theta_{-i}) > \hat{u}_i\big(f(\tilde{\theta}_i, \theta'_{-i}), \theta_{-i}\big),$$

for all $\theta_{-i} \in \hat{M}_k$ and $u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > u_i\big(\bar{m}'_i, (\tilde{\theta}_i, \theta'_{-i})\big)$.

Define $y(\theta'_{-i}) \in \Delta(A)$ such that $y(\theta'_{-i}) = \bar{m}'_i$. Then we have $u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) \geq u_i\big(y(\theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)$ and

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

for all $\theta_{-i} \in \Theta_{-i}$ such that $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

Now let $y : \Theta_{-i} \to \Delta(A)$ be such that $y(\theta'_{-i})$ is as defined above for each $\theta'_{-i} \in \Theta_{-i}$. Then $y \in Y_i^w[\tilde{\theta}_i]$ and it satisfies the inequality in the preference-reversal condition. □

Proof of Lemma 6.5. ($\Rightarrow$) Suppose the SCF f satisfies the preference-reversal condition. We argue that f satisfies the sign-preserving property.

Pick an unacceptable deception β. Then, according to the preference-reversal condition, 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 $\tilde{\theta}_i \sim_i^f \theta'_i$, there

exists $y \in Y_i^w[\tilde{\theta}_i]$ such that

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

for all $\theta_{-i} \in \Theta_{-i}$ and $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$

Pick any $\tilde{\theta}_i \sim_i^f \theta'_i$ and $y \in Y_i^w[\tilde{\theta}_i]$ that satisfies the above condition. Fix $\theta_{-i} \in \Theta_{-i}$ and $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$. Since $y \in Y_i^w[\tilde{\theta}_i]$, we have $u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) \geq u_i\big(y(\theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)$. Also, since $\theta'_{-i} \in \beta_{-i}(\theta'_{-i})$, the preference-reversal condition implies that

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

Then the above three inequalities imply that

$$\big(v_i(\tilde{\theta}_i, \theta'_{-i}) - v_i(\theta_i, \theta'_{-i})\big)\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{y(\theta'_{-i})}\big) > 0 \text{ and } \big(v_i(\tilde{\theta}_i, \theta'_{-i}) - v_i(\theta_i, \theta_{-i})\big)\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{y(\theta'_{-i})}\big) > 0.$$

Therefore, it must be that $\text{sign}\Big(v_i(\theta_i, \theta'_{-i}) - v_i(\tilde{\theta}_i, \theta'_{-i})\Big) = \text{sign}\Big(v_i(\theta_i, \theta_{-i}) - v_i(\tilde{\theta}_i, \theta'_{-i})\Big) \neq 0$.

($\Leftarrow$) Suppose the SCF f has interior transfers, and satisfies semi-strict EPIC and the sign-preserving property. We argue that f satisfies the preference-reversal condition.

Pick an unacceptable deception β. Then, the sign-preserving property implies that 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 $\tilde{\theta}_i \sim_i^f \theta'_i$

$$\text{sign}\Big(v_i(\theta_i, \theta'_{-i}) - v_i(\tilde{\theta}_i, \theta'_{-i})\Big) = \text{sign}\Big(v_i(\theta_i, \theta_{-i}) - v_i(\tilde{\theta}_i, \theta'_{-i})\Big) \neq 0,$$

for all $\theta_{-i} \in \Theta_{-i}$ and $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

Pick any $\tilde{\theta}_i \sim_i^f \theta'_i$. Fix $\theta'_{-i} \in \Theta_{-i}$. Since $\theta'_{-i} \in \beta_{-i}(\theta'_{-i})$, the sign-preserving property implies that $v_i(\theta_i, \theta'_{-i}) \neq v_i(\tilde{\theta}_i, \theta'_{-i})$. Suppose $v_i(\theta_i, \theta'_{-i}) < v_i(\tilde{\theta}_i, \theta'_{-i})$. (The argument for the case $v_i(\theta_i, \theta'_{-i}) > v_i(\tilde{\theta}_i, \theta'_{-i})$ is similar and left to the reader.)

As $\tilde{\theta}_i \sim_i^f \theta'_i$ and $\theta'_i \not\sim_i^f \theta_i$, we have $\tilde{\theta}_i \not\sim_i^f \theta_i$. By semi-strict EPIC, $u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > u_i\big(f(\theta_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)$ and $u_i\big(f(\theta_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big) > u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big)$. Therefore,

$$v_i(\tilde{\theta}_i, \theta'_{-i})\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{f(\theta_i, \theta'_{-i})}\big) > \tau_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \tau_i^{f(\theta_i, \theta'_{-i})} > v_i(\theta_i, \theta'_{-i})\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{f(\theta_i, \theta'_{-i})}\big).$$

Then, since $v_i(\tilde{\theta}_i, \theta'_{-i}) > v_i(\theta_i, \theta'_{-i})$, it follows that $q_i^{f(\tilde{\theta}_i, \theta'_{-i})} > q_i^{f(\theta_i, \theta'_{-i})}$.

Let

$$\hat{\theta}_{-i} \in \underset{\theta_{-i} \in \Theta_{-i}:\, \theta'_{-i} \in \beta_{-i}(\theta_{-i})}{\arg\max} v_i(\theta_i, \theta_{-i}).$$

Since $v_i(\theta_i, \theta'_{-i}) < v_i(\tilde{\theta}_i, \theta'_{-i})$, by the sign-preserving property, we have $v_i(\theta_i, \hat{\theta}_{-i}) < v_i(\tilde{\theta}_i, \theta'_{-i})$.

Pick any δ such that $v_i(\theta_i, \hat{\theta}_{-i}) < \delta < v_i(\tilde{\theta}_i, \theta'_{-i})$. Since f has interior transfers, it follows that $\tau_i^{f(\tilde{\theta}_i, \theta'_{-i})} \in (-z, z)$. Hence, we can find a sufficiently small $\epsilon > 0$, and define τ_i^ϵ such that $\tau_i^\epsilon \equiv \tau_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \epsilon\delta\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{f(\theta_i, \theta'_{-i})}\big) \in (-z, z)$.

Consider the lottery $\ell^\epsilon \equiv (1-\epsilon) f(\tilde{\theta}_i, \theta'_{-i}) + \epsilon f(\theta_i, \theta'_{-i})$. Now, define the lottery $y(\theta'_{-i}) \in \Delta(\bar{A})$ as follows: For all $(q, \tau_i, \tau_{-i}) \in \bar{A}$, let

$$y(\theta'_{-i})[(q, \tau_i, \tau_{-i})] = \begin{cases} 0, & \text{if } \tau_i \neq \tau_i^\epsilon, \\ \sum_{\tau'_i \in [-z,z]} \ell^\epsilon\big[\big(q, (\tau'_i, \tau_{-i})\big)\big], & \text{if } \tau_i = \tau_i^\epsilon. \end{cases}$$

By construction, we have that $q_i^{y(\theta'_{-i})} = (1-\epsilon) q_i^{f(\tilde{\theta}_i, \theta'_{-i})} + \epsilon q_i^{f(\theta_i, \theta'_{-i})}$ and $\tau_i^{y(\theta'_{-i})} = \tau_i^\epsilon$.

Note that $\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{y(\theta'_{-i})}\big) = \epsilon\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{f(\theta_i, \theta'_{-i})}\big)$, which is positive because $q_i^{f(\tilde{\theta}_i, \theta'_{-i})} > q_i^{f(\theta_i, \theta'_{-i})}$. By the definition of δ and $\tau_i^{y(\theta'_{-i})}$, we obtain the following:

$$\begin{aligned} v_i(\tilde{\theta}_i, \theta'_{-i})\epsilon\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{f(\theta_i, \theta'_{-i})}\big) > \delta\epsilon\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{f(\theta_i, \theta'_{-i})}\big) > v_i(\theta_i, \hat{\theta}_{-i})\epsilon\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{f(\theta_i, \theta'_{-i})}\big) \\ \Rightarrow v_i(\tilde{\theta}_i, \theta'_{-i})\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{y(\theta'_{-i})}\big) > \tau_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \tau_i^{y(\theta'_{-i})} > v_i(\theta_i, \hat{\theta}_{-i})\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{y(\theta'_{-i})}\big). \end{aligned}$$

Since the above inequalities are strict, it is without loss of generality to assume that $y(\theta'_{-i}) \in \Delta(A)$.[22] Then, the first inequality implies that $u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) \geq u_i\big(y(\theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)$. By the very definition of $\hat{\theta}_{-i}$, the second inequality implies that

$$\tau_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \tau_i^{y(\theta'_{-i})} > v_i(\theta_i, \hat{\theta}_{-i})\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{y(\theta'_{-i})}\big) \geq v_i(\theta_i, \theta_{-i})\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^{y(\theta'_{-i})}\big),$$

for all $\theta_{-i} \in \Theta_{-i}$ such that $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$. Thus, $u_i\big(y(\theta'_{-i}), (\theta_i, \theta_{-i})\big) > u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big)$, for all $\theta_{-i} \in \Theta_{-i}$ such that $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

Let $y : \Theta_{-i} \to \Delta(A)$ be such that $y(\theta'_{-i})$ is defined as above for all $\theta'_{-i} \in \Theta_{-i}$. Clearly, $y \in Y_i^w[\tilde{\theta}_i]$ and it satisfies the inequality in the preference-reversal condition. □

Proof of Lemma 6.7. Suppose $\bar{Q}$ is rich. Let f be a deterministic and interior SCF. Pick an unacceptable deception β. Then, the sign-preserving property implies that 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 $\tilde{\theta}_i \sim_i^f \theta'_i$

$$\text{sign}\Big(w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big) - w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)\Big)$$

[22] As the set of probability measures with finite support is dense in $\Delta(\bar{A})$, we can find a lottery ℓ with a finite support that is close enough to $y(\theta'_{-i})$ such that the above two inequalities are satisfied when we replace $(q_i^{y(\theta'_{-i})}, \tau_i^{y(\theta'_{-i})})$ by (q_i^ℓ, τ_i^ℓ). Next, since $\bar{A}$ is the closure of A and ℓ has a finite support, we can approximate the points in the support of ℓ by points in A to obtain a lottery in $\Delta(A)$ that satisfies the two inequalities.

$$= \text{sign}\Big(w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big) - w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)\Big)$$
$$\neq 0,$$

for all $\theta_{-i} \in \Theta_{-i}$ and $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

Pick any $\tilde{\theta}_i \sim_i^f \theta'_i$. Fix $\theta'_{-i} \in \Theta_{-i}$. It thus follows from the sign-preserving property that $w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big) \neq w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)$. Suppose $w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big)$. (The argument for the other case is similar and left to the reader.)

For any $\epsilon > 0$, let $q_i^\epsilon = q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \epsilon$. Then, by the definition of partial derivative,

$$\lim_{\epsilon \to 0} \frac{v_i\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})}, (\tilde{\theta}_i, \theta'_{-i})\big) - v_i\big(q_i^\epsilon, (\tilde{\theta}_i, \theta'_{-i})\big)}{q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^\epsilon} = w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) \text{ and}$$
$$\lim_{\epsilon \to 0} \frac{v_i\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})}, (\theta_i, \theta_{-i})\big) - v_i\big(q_i^\epsilon, (\theta_i, \theta_{-i})\big)}{q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^\epsilon} = w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big),$$

for all $\theta_{-i} \in \Theta_{-i}$ such that $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

Let

$$\hat{\theta}_{-i} \in \underset{\theta_{-i} \in \Theta_{-i}:\, \theta'_{-i} \in \beta_{-i}(\theta_{-i})}{\arg\max} w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big).$$

Since $w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta'_{-i})\big)$, by the sign-preserving property, we have that $w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \hat{\theta}_{-i})\big)$.

Pick any δ such that $w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) > \delta > w_i\big((\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \hat{\theta}_{-i})\big)$. Since Θ_{-i} is finite, by the very definition of $\hat{\theta}_{-i}$, there exists $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*)$, we have

$$\frac{v_i\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})}, (\tilde{\theta}_i, \theta'_{-i})\big) - v_i\big(q_i^\epsilon, (\tilde{\theta}_i, \theta'_{-i})\big)}{q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^\epsilon} > \delta > \frac{v_i\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})}, (\theta_i, \theta_{-i})\big) - v_i\big(q_i^\epsilon, (\theta_i, \theta_{-i})\big)}{q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^\epsilon},$$

for all $\theta_{-i} \in \Theta_{-i}$ such that $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

For all $\epsilon \in (0, \epsilon^*)$, define $\tau_i^\epsilon = \tau_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \delta(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^\epsilon)$. Then, by construction of δ and τ_i^ϵ, we obtain

$$\begin{aligned} v_i\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})}, (\tilde{\theta}_i, \theta'_{-i})\big) - v_i\big(q_i^\epsilon, (\tilde{\theta}_i, \theta'_{-i})\big) &> \tau_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \tau_i^\epsilon \\ &> v_i\big(q_i^{f(\tilde{\theta}_i, \theta'_{-i})}, (\theta_i, \theta_{-i})\big) - v_i\big(q_i^\epsilon, (\theta_i, \theta_{-i})\big), \end{aligned}$$

for all $\theta_{-i} \in \Theta_{-i}$ such that $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

Recall that $q_i^\epsilon = q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \epsilon < q_i^{f(\tilde{\theta}_i, \theta'_{-i})}$ and $\tau_i^\epsilon = \tau_i^{f(\tilde{\theta}_i, \theta'_{-i})} - \delta(q_i^{f(\tilde{\theta}_i, \theta'_{-i})} - q_i^\epsilon) < \tau_i^{f(\tilde{\theta}_i, \theta'_{-i})}$.

Since f is in the interior, there exist $\hat{\epsilon} \in (0, \epsilon^*)$ and $y(\theta'_{-i}) \in A$ such that $q_i^{y(\theta'_{-i})} = q_i^{\hat{\epsilon}}$ and $\tau_i^{y(\theta'_{-i})} = \tau_i^{\hat{\epsilon}}$. Then,

$$v_i\big(q_i^{f(\tilde{\theta}_i,\theta'_{-i})}, (\tilde{\theta}_i, \theta'_{-i})\big) - v_i\big(q_i^{y(\theta'_{-i})}, (\tilde{\theta}_i, \theta'_{-i})\big) > \tau_i^{f(\tilde{\theta}_i,\theta'_{-i})} - \tau_i^{y(\theta'_{-i})}$$
$$> v_i\big(q_i^{f(\tilde{\theta}_i,\theta'_{-i})}, (\theta_i, \theta_{-i})\big) - v_i\big(q_i^{y(\theta'_{-i})}, (\theta_i, \theta_{-i})\big),$$

for all $\theta_{-i} \in \Theta_{-i}$ such that $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

As $y(\theta'_{-i}) \in A$, the first inequality implies that $u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big) \geq u_i\big(y(\theta'_{-i}), (\tilde{\theta}_i, \theta'_{-i})\big)$. The second inequality implies that $u_i\big(y(\theta'_{-i}), (\theta_i, \theta_{-i})\big) > u_i\big(f(\tilde{\theta}_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big)$, for all $\theta_{-i} \in \Theta_{-i}$ such that $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$.

Let $y : \Theta_{-i} \to A$ be such that $y(\theta'_{-i})$ is defined as above for all $\theta'_{-i} \in \Theta_{-i}$. Clearly, $y \in Y_i^w[\tilde{\theta}_i]$ and it satisfies the inequality in the preference-reversal condition. □

References

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

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

[3] Artemov, G., T. Kunimoto, R. Serrano, "Robust Virtual Implementation: Toward a Reinterpretation of the Wilson Doctrine," *Journal of Economic Theory*, vol. 148, (2013), 424-447.

[4] Battigalli, P., and M. Siniscalchi, "Rationalization and Incomplete Information", *Advances in Theoretical Economics*, vol. 3, no. 1, (2003).

[5] Bergemann, D., and S. Morris, "Robust Implementation: The Role of Large Type Spaces," *Working Paper*, (2005a).

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

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

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

[9] Bergemann, D., and S. Morris, "Robust Virtual Implementation," *Theoretical Economics*, vol. 4, (2009b), 45-88.

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

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

[12] Bergemann, D., and S. Morris, "Belief-Free Rationalizability and Informational Robustness," *Games and Economic Behavior*, vol. 104, (2017), 744-759.

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

[14] Chen, Y-C, T. Kunimoto, Y. Sun, and S. Xiong, "Maskin Meets Abreu and Matsushima," *Theoretical Economics*, vol. 17, (2022), 1683-1717.

[15] Chen, Y-C, T. Kunimoto, Y. Sun, and S. Xiong, "Rationalizable Implementation in Finite Mechanisms," *Games and Economic Behavior*, vol. 129, (2021), 181-197.

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

[17] Guo, H., and N.C. Yannelis, "Robust Coalitional Implementation," *Games and Economic Behavior*, 132, (2022), 553-575.

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

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

[20] Jain, R., M. Lombardi, and C. Müller, "An alternative equivalent formulation for robust implementation," *Games and Economic Behavior*, vol. 142, (2023), 368-380.

[21] Kunimoto, T. and R. Saran, "Robust Implementation in Rationalizable Strategies in General Mechanisms," *Working Paper*, (2020), Singapore Management University.

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

[23] Kunimoto, T., R. Saran, and R. Serrano, "Interim Rationalizable Implementation of Functions," *Mathematics of Operations Research*, (2023), forthcoming.

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

[25] Müeller, C., "Robust Implementation in Weakly Perfect Bayesian Strategies," *Journal of Economic Theory*, 189, (2020), 105038.

[26] Ollár, M. and A. Penta, "Full Implementation and Belief Restrictions," *American Economic Review*, vol. 107, (2017), 2243-2377.

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

[28] Serrano, R., and R. Vohra, "A characterization of Virtual Bayesian Implementation." *Games and Economic Behavior*, vol. 50, (2005), 312-331.

[29] Xiong, S., "Rationalizable Implementation of Social Choice Functions: Complete Characterization," *Theoretical Economics*, vol. 18, (2023), 197-230.