Robust Implementation in Weakly Perfect Bayesian Strategies*

Christoph Müller†
University of Queensland

March 2, 2018

Abstract

We examine a strong form of robust implementation in dynamic mechanisms that is both belief- and belief-revision-free. Specifically, we characterize robust wPBE-implementation, that is, full implementation in weak Perfect Bayesian equilibrium across all type spaces. We introduce a dynamic robust monotonicity condition that is weaker than Bergemann and Morris' (2011) robust monotonicity condition and show that under a conditional no total indifference condition, ex-post incentive compatibility and dynamic robust monotonicity characterize robust wPBE-implementation in general dynamic mechanisms. We also introduce a notion of weakly rationalizable implementation (wr-implementation) and prove that it is equivalent to robust wPBE-implementation. Applied to static mechanisms, wr-implementation *exactly* characterizes a version of static belief-free implementation.

KEYWORDS: robust implementation, dynamic mechanisms, weak rationalizability, weak perfect Bayesian equilibrium, dynamic robust monotonicity.

*This paper extends and supersedes parts of an earlier working paper circulated under the title "Robust Implementation in Weakly Rationalizable Strategies." I thank everyone who commented on that paper, and Andy McLennan and Claudio Mezzetti for comments on the current paper. All errors are my own.

†Contact: c.mueller@uq.edu.au

1 Introduction

Since Bergemann and Morris' (2005; 2009a; 2011) seminal series of papers, there has been a renewed and growing interest in implementing social choice functions in a robust way. In their work Bergemann and Morris equate robustness with belief-freeness, that is, independence of the agents' beliefs and higher order beliefs about the state of the world. In addition to robustness, it is advantageous to achieve full implementation, that is, to not only guarantee that the implementing mechanism contains acceptable equilibria but also to rule out that there are any unacceptable ones. In short, it is desirable to achieve full robust implementation.

Full robust implementation, however, is demanding. For this reason, researchers have searched for more permissive results by weakening the robustness requirement or moving to notions of approximate full implementation or both (e.g. Artemov et al., 2013; Müller, 2016; Ollár and Penta, 2017; see also Bergemann and Morris, 2009b). In this search, the complementary approach of enlarging the class of mechanisms to dynamic mechanisms (roughly, extensive game forms) is only beginning to be investigated (e.g. Müller, 2016; see also Penta, 2015)[1]. While still largely unexplored in robust implementation, dynamic mechanisms have been successfully employed in full non-robust implementation, both in complete information environments (e.g. Moore and Repullo, 1988; Abreu and Sen, 1990; Vartiainen, 2007) and in the case of incomplete information on which we focus in this paper (e.g. Brusco, 1995, 2006; Bergin and Sen, 1998; Duggan, 1998; Baliga, 1999). This makes dynamic mechanisms a promising alternative to weakening the robustness notion or abandoning exact in favor of approximate full implementation.

An intrinsic complication with dynamic mechanisms, however, is that they may confront agents with information sets that these agents expected with probability zero. At such "surprise" information sets agents cannot Bayesian update their previous beliefs and instead have to revise them in other ways (since we maintain throughout that agents use Bayesian updating when not surprised, from now on, by "belief revision" we in particular refer to the belief revision at surprise information sets). Many of the positive results for dynamic non-robust implementation are derived under specific assumptions on the belief revision.[2] Importantly, the existing positive result for dynamic robust implementation in Müller (2016) not only applies to a concept of approximate implementation, but crucially also depends on a specific belief-revision assumption. While belief-revision assumptions in various forms are widely re-

[1] We discuss these and other papers on implementation by dynamic mechanisms in Subsection 5.1.

[2] Assumptions on the belief revision at surprise information sets are often implicit in the solution concept, for example in perfect Bayesian equilibrium (used by Brusco, 1995, 2006) and sequential equilibrium (used by Baliga, 1999, and in effect also by Bergin and Sen, 1998). Such assumptions are even common under complete information, e.g. in subgame perfect equilibrium (used by Moore and Repullo, 1988; Abreu and Sen, 1990; Vartiainen, 2007). A notable exception is Duggan (1998), who obtained positive results in a more specialized setting without such belief-revision assumptions. For more details, see Subsection 5.1.

lied on in the literature, mechanism designers may, however, not always be confident which assumptions (if any) are valid in practice. Accordingly, there are not one but two robustness concerns in dynamic mechanisms. First, as in static mechanisms, achieving robustness with respect to the agents' initial beliefs about the state of the world, expressed by conditions such as Bergemann and Morris' belief-freeness. Second, achieving robustness with respect to the agents' belief revision processes.

In short, the state of the literature raises several questions. How much can dynamic mechanisms weaken the necessary conditions for full robust implementation, particularly if one insists on belief-freeness and exact implementation? And to what extent does the success of dynamic mechanisms in full robust implementation depend on belief-revision assumptions? In this paper we take a step towards answering these questions by adopting weakly perfect Bayesian equilibrium (wPBE) and studying full robust implementation in wPBE (robust wPBE-implementation). Robust wPBE-implementation is a notion of belief-free and exact implementation by dynamic mechanisms, and like wPBE maintains Bayesian updating but does not make any assumptions on the belief revision at surprise information sets. Specifically, first, we introduce and then provide necessary and sufficient conditions for a notion of weakly rationalizable implementation (wr-implementation). Second, we prove that wr-implementation is equivalent to robust wPBE-implementation (Theorem 2). In combination, Theorems 1 and 2 characterize robust wPBE-implementation, and thus provide one answer to the first question from above and establish an important benchmark for the second.

In more detail, we believe that a characterization of robust wPBE-implementation contributes in several ways. First, our sufficiency result (Theorem 1(b) in conjunction with Theorem 2) provides conditions under which an arguably very strong form of robust implementation can be guaranteed. The concern of the robust mechanism design literature is that if a mechanism designer makes a mistake in modeling the agents' belief hierarchies about the state of the world, then a standard (i.e. non-robust) mechanism might "malfunction" and not implement the desired social choice function. This concern is real, as even small changes in the agents' belief hierarchies can alter equilibrium outcomes (see e.g. Rubinstein, 1989, Weinstein and Yildiz, 2007, and Penta, 2013, for static games and Penta, 2012, for dynamic games). Following the foundational contributions of Bergemann and Morris (2005) and Chung and Ely (2007), the robust mechanism design literature identifies mechanisms that do not depend on "details" about the agents' belief hierarchies. With belief-freeness, robust wPBE-implementation maintains the strongest and thus most desirable static robustness condition of this literature, namely independence from the initial belief hierarchies about the state of the world, while extending the analysis to dynamic mechanisms. With respect to the dynamic aspects, robust wPBE-implementation is belief-revision-free in the sense that it only imposes the rather minimal assumptions of sequential rationality (generalizing rationality from the static

case) and Bayesian updating. Therefore, robust wPBE-implementation provides "maximal" robustness with respect to both initial and revised beliefs, requiring no knowledge of either.

Second, maintaining belief-freeness has the advantage that our sufficient conditions are easily comparable to those for robust implementation by static mechanisms derived by Bergemann and Morris (2011) (henceforth, BM), giving us a sense of what dynamic mechanisms add. Clearly, since they guarantee strong dynamic robustness, we cannot expect the conditions for robust wPBE-implementation to be overly permissive. But due to admitting dynamic mechanisms and despite the absence of belief-revision assumptions, robust wPBE-implementation is nonetheless more permissive than BM's robust implementation by static mechanisms.

We could relax our sufficient conditions further by imposing assumptions on the agents' belief revision or initial beliefs or both. However, without understanding robust wPBE-implementation, we would not know when such stronger assumptions are actually necessary. In other words, third, our necessary conditions for robust wPBE-implementation (Theorem 1(a) in conjunction with Theorem 2) delineate an important boundary to robust implementation that can only be overcome at the cost of such stronger assumptions.

While extreme, insisting on belief-revision-freeness means that robust wPBE-implementation can build a basis for disentangling the effect of simply permitting dynamic mechanisms from simultaneously making belief-revision assumptions. Like most past results for non-robust implementation, potential future results for robust implementation may rely on particular belief-revision assumptions. Being able to compare such potential future results to our necessary conditions is valuable, as it can inform mechanism designers who have to manage the trade-off between avoiding such belief-revision assumptions (which they may not be completely confident in) and being able to implement additional social choice functions.

Fourth, an advantage of our weak informational assumptions is that they allow us to completely characterize implementability in fairly general environments: under a conditional no-total indifference condition, our necessary and sufficient conditions (for wr- and also for robust wPBE-implementability) coincide. Beyond assuming finite environments, our analysis relies on the agents maximizing expected utility and on mechanisms being able to terminate in lotteries over pure outcomes. As we discuss in Subsection 5.1, to the best of our knowledge, a tight characterization of implementability by general dynamic mechanisms is a first in the literature on full (robust and also non-robust) implementation in incomplete information environments. Therefore, our results may also contribute to the understanding of dynamic implementation more generally.

As already mentioned, we do not directly characterize robust wPBE-implementation, but instead establish it as equivalent to wr-implementation (Theorem 2). In this respect, our approach is analogous to BM's, who characterize (static) robust implementation by relating it to a notion of rationalizable implementation. A benefit of this approach is that, unlike

robust wPBE-implementation, wr-implementation is defined and can be analyzed without introducing the machinery of type spaces, which can be an important simplifying factor in practice. In other words, mechanism designers may prefer to work with wr-implementation instead of robust wPBE-implementation.

Let us make three remarks about Theorem 2. First, because it establishes a tight equivalence between wr-implementation and robust wPBE-implementation, Theorem 2 confirms that our definition of wr-implementation is "correct" for our purposes. Wr-implementation is based on a slight variation of Battigalli's (1999, 2003) weak rationalizability, and includes a condition that Theorem 2 shows to guarantee the existence of a wPBE on all type spaces. Second, an advantage of Theorem 2 is that it applies to a very broad class of general dynamic mechanisms, an aspect in which it goes beyond its game-theoretic counterpart by Battigalli (1999). For example, Theorem 2 applies as long as we rule out mechanisms that violate a measurability condition. Finally, Theorem 2 has implications for static robust implementation. To elaborate from above, BM already proposed a notion of rationalizable implementation by static mechanisms, and showed that it is *almost* equivalent to (static) robust implementation. Applying Theorem 2 to static mechanisms reveals that the notion of wr-implementation by static mechanisms modifies BM's rationalizable implementation in a way that results in an *exact* equivalence to (static) robust implementation (see Subsection 4.4).

Given the equivalence between robust wPBE-implementation and wr-implementation we simply focus on characterizing the latter (Theorem 1). In Subsection 3.1, we propose the comparatively simple notion of robust preference reversals. Using this notion, we define dynamic robust monotonicity (dr-monotonicity), and show that dr-monotonicity and ex-post incentive compatibility (epIC) are necessary and, together with a conditional no total indifference condition, also sufficient for wr-implementation. Dr-monotonicity is related to but weaker than BM's robust monotonicity condition, and epIC is weaker than semi-strict epIC. Robust monotonicity and semi-strict epIC are necessary for BM's rationalizable implementation. Additional value from Theorem 1 derives because wr-implementation is also of independent interest beyond being a proxy for robust wPBE-implementation (see for example Subsection 5.2).

As a benefit, our proof of Theorem 1 shows that in order to wr-implement social choice functions, it suffices to restrict attention to countable mechanisms, that is, mechanisms with countable strategy sets. Restricting attention to countable mechanisms is convenient as it avoids dealing with the technicalities of uncountable mechanisms. Specifically, our sufficiency proof shows that every wr-implementable social choice function can in fact be wr-implemented by a countable mechanism. And our necessary conditions apply to general dynamic mechanisms. Consequently, wr-implementation by countable and by general mechanisms are equivalent, and focusing on the former entails no loss.

Organization of the Paper. Section 2 introduces the notation, the environment, countable mechanisms and wr-implementation. Section 3 derives necessary and sufficient conditions for wr-implementation. Section 4 relates wr-implementation to robust wPBE-implementation and introduces general mechanisms. Section 5 discusses some related literature and concludes. Appendix A comments on our definition of wr-implementation, and Appendix B collects the proofs omitted from the main text.

2 Preliminaries

Let $\mathcal{J} = \{1, \ldots, J\}$ be a finite set. If $(Z_j)_{j \in \mathcal{J}}$ is a family of sets Z_j indexed by $\mathcal{J}$, then Z denotes the Cartesian product $\prod_{j \in \mathcal{J}} Z_j$. If $z_j \in Z_j$ for all $j \in \mathcal{J}$, then z denotes $(z_1, \ldots, z_J)$ and, for each $i \in \mathcal{J}$, $z_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_J)$. At times we ignore the correct order of tuples and write (z_j, z_{-j}) for z and, if for all $j \in \mathcal{J}$, $(v_j, w_j) \in Z_j = V_j \times W_j$ for some sets V_j and W_j, $((v_1, \ldots, v_J), (w_1, \ldots, w_J))$ for the element $(v_j, w_j)_{j \in \mathcal{J}}$ of Z. If V and Z are sets, then 2^Z denotes the power set of Z and Z^V the set of functions mapping V to Z.

We endow every countable set with its discrete topology, every product of topological spaces with the product topology, and every topological space with its Borel σ-algebra.[3] If $(Z, \mathcal{Z})$ is a measurable space then $\Delta(Z)$ denotes the set of probability measures on Z. If in addition $\{z\} \in \mathcal{Z}$, then $\delta(z)$ denotes the degenerate probability measure on Z that places probability one on $z \in Z$. If Z_j is endowed with a topology for all $j \in \mathcal{J}$ and $p \in \Delta(\prod Z_j)$, then $marg_{Z_j} p$ denotes the j-th coordinate marginal of p, that is, the measure induced from p by the projection from $\prod Z_j$ to Z_j. If Z is countable and $p \in \Delta(Z)$, then $supp(p) = \{z \in Z : p(z) > 0\}$ denotes the support of p.

We let Ord denote the class of ordinal numbers.

2.1 Environment

There is a finite set $\mathcal{I} = \{1, \ldots, I\}$ of at least two agents. Every agent $i \in \mathcal{I}$ has a nonempty and finite payoff type space Θ_i and privately observes a payoff type $\theta_i \in \Theta_i$ that represents her payoff-relevant information. There is a nonempty and finite set X of pure outcomes. We let $Y = \Delta(X)$, endow Y with the relative Euclidean topology and call elements $y \in Y$ lotteries (over X) or outcomes.

The agents have expected utility preferences over outcomes which are interdependent in that i's utility can depend on the payoff-relevant information θ_{-i} of her opponents. Specifically, we let $u_i(x, \theta)$ denote the von Neumann-Morgenstern utility that $i \in \mathcal{I}$ derives from the pure outcome x if the payoff type profile is $\theta \in \Theta$, and, in a slight abuse of notation, $u_i(y, \theta)$ the expected utility that i derives from lottery y if the payoff type profile is θ.

[3] Our results also hold under alternative assumptions; see Footnotes 16 and 18 for details.

Following the belief-free paradigm, i's (initial) belief $\psi_i \in \Delta(\Theta_{-i})$ about $-i$'s payoff type profile θ_{-i} is not part of our description of an environment. Instead, we will work with implementation concepts that consider all possible beliefs ψ_i. Hence we also do not take expectations of i's utility with respect to beliefs ψ_i here; rather, such expectations will implicitly appear when we define sequential rationality in Subsection 2.3. Still, we already note that in another abuse of notation we write $\theta_j \in supp(\psi_i)$ if $\psi_i \in \Delta(\Theta_{-i})$, $j \neq i$ and $\theta_j \in \Theta_j$ is in the support of $marg_{\Theta_j}\psi_i$.

2.2 Countable Mechanisms

Our characterization of wr-implementation conveniently applies to both a general class of mechanisms and a class of simpler mechanisms that we will call countable mechanisms: our necessary conditions apply even if the designer has general mechanisms at her disposal, and our sufficiency proof demonstrates that every wr-implementable social choice function is indeed wr-implementable by a countable mechanism.

Considering general mechanisms requires the careful treatment of measurability issues. Purely to simplify the exposition, we therefore limit attention to countable mechanisms until we reach Subsection 4.2 (even though for brevity we will not write the qualifier "countable" every time we write "mechanism", even though we already adopt a notation that easily generalizes later and even though we immediately formulate those proofs that we relegate to Appendix B for general mechanisms). Moreover, we will only summarize important features of countable mechanisms here. Subsection 4.2 will formally define countable and general mechanisms, and present the appropriate generalizations of the other definitions that we make in the remainder of this Section 2.

A countable (dynamic) mechanism $\Gamma = \langle A, H, (\mathbb{H}_i)_{i\in\mathcal{I}}, P, C\rangle$ is an extensive game form with perfect recall and no trivial decisions that conforms to the following description. Its first component, A, is a countable set of actions. Its second component, H, is a set of histories $h = (a_1, \ldots, a_n)$, which are finite sequences of actions. Play starts at the (typically non-terminal) initial history $\emptyset$. At each non-terminal history $h = (a_1, \ldots, a_n)$, the agent $P(h)$ specified by the player function P chooses an action from the set $A(h) = \{a \in A : (h, a) \in H\}$. Here, (h, a) denotes the history $(a_1, \ldots, a_n, a)$. Once a history h such that $A(h) = \emptyset$ — that is, a terminal history — is reached, the mechanism concludes and the lottery $C(h) \in Y$ specified by the outcome function C obtains. The set $\mathbb{H}_i$ partitions the set of all histories at which i moves into information sets $\mathcal{H}$. Whenever i moves she knows the information set, but not the history she is at. A strategy s_i for player specifies an (available) action for each information set $\mathcal{H} \in \mathbb{H}_i$. Letting S_i denote the set of i's strategies, a countable mechanism is such that S_i is *countable* for all $i \in \mathcal{I}$.

The terminal history induced by strategy profile $s \in S$ is denoted by $\zeta(s)$. We use the

symbol $\preceq$ to indicate precedence among histories, and also to indicate precedence among i's information sets. We let $S_i(\mathcal{H}) = \{s_i \in S_i : \exists s_{-i} \in S_{-i} \exists h \in \mathcal{H}, h \preceq \zeta(s)\}$ be the set of i's strategies that *admit* j's information set $\mathcal{H}$, $j \in \mathcal{I}$, and $S_{-i}(\mathcal{H})$ be the set of strategy profiles of $-i$ admitting $\mathcal{H}$. For each $\mathcal{J} \subseteq \mathcal{I}$, we let

$$\mathbb{H}_i((s_j)_{j\in\mathcal{J}}) = \left\{\mathcal{H} \in \mathbb{H}_i : \left(\exists h \in \mathcal{H}, (s_j)_{j\in\mathcal{I}\setminus\mathcal{J}} \in \prod_{j\in\mathcal{J}\setminus\mathcal{I}} S_j\right)\left(h \preceq \zeta(s)\right)\right\}$$

denote the set of i's information sets admitted by $(s_j)_{j\in\mathcal{J}}$. For $A \subseteq S$, $\mathbb{H}_i(A)$ denotes the union of sets $\mathbb{H}_i(s)$, where $s \in A$. Moreover, $\Sigma_i = S_i \times \Theta_i$, $\Sigma_{-i} = S_{-i} \times \Theta_{-i}$ and $\Sigma_{-i}(\mathcal{H}) = S_{-i}(\mathcal{H}) \times \Theta_{-i}$.

A countable static mechanism consists of a countable strategy set S_i for each $i \in \mathcal{I}$ and an outcome function $C : S \to Y$. Formally, a countable static mechanism is simply a special countable (dynamic) mechanism (see Subsection 4.2).

2.3 Beliefs and Sequential Rationality in Countable Mechanisms

Given a mechanism, player i has beliefs about her opponents' strategies and payoff types. These beliefs are captured by an indexed family of probability measures on $(\Sigma_{-i}, \mathcal{B}_{-i})$, where $\mathcal{B}_{-i} = 2^{\Sigma_{-i}}$ denotes the discrete σ-algebra on Σ_{-i}. Each measure represents i's belief at a different point in the mechanism. Precisely, the index set is $\bar{\mathbb{H}}_i = \mathbb{H}_i \cup \{\{\emptyset\}\}$ and i has a belief on $(\Sigma_{-i}, \mathcal{B}_{-i})$ at each of her information sets and at the initial history (even if the initial history does not comprises one of her information sets). Player i's beliefs so indexed form a conditional probability system.

Definition 1 (Rényi, 1955) *A conditional probability system (CPS) on $(\Sigma_{-i}, \bar{\mathbb{H}}_i)$ is a function $\mu_i : \mathcal{B}_{-i} \times \bar{\mathbb{H}}_i \to [0,1]$ such that*

(a) for all $\mathcal{H} \in \bar{\mathbb{H}}_i$, $\mu_i(\cdot|\mathcal{H})$ is a probability measure on $(\Sigma_{-i}, \mathcal{B}_{-i})$.

(b) for all $\mathcal{H} \in \bar{\mathbb{H}}_i$, $\mu_i(\Sigma_{-i}(\mathcal{H})|\mathcal{H}) = 1$.

(c) for all $\mathcal{H}, \mathcal{H}' \in \bar{\mathbb{H}}_i$ and $D \in \mathcal{B}_{-i}$ such that $D \subseteq \Sigma_{-i}(\mathcal{H})$, if $\mathcal{H}' \preceq \mathcal{H}$ then

$$\mu_i(D|\mathcal{H})\mu_i(\Sigma_{-i}(\mathcal{H})|\mathcal{H}') = \mu_i(D|\mathcal{H}').$$

In particular, by Condition (c), "i uses Bayesian updating whenever applicable," that is, i's beliefs are related by the rules of conditional probability whenever possible. However, Condition (c) imposes no assumption on the belief revision if an information set surprises i. An information set $\mathcal{H}$ is a *surprise* to i if at the immediate predecessor of $\mathcal{H}$, i places probability zero on $\Sigma_{-i}(\mathcal{H})$ and thus is certain that $\mathcal{H}$ will not be reached.

We let $\Delta(\Sigma_{-i})$ denote the set of probability measures on $(\Sigma_{-i}, \mathcal{B}_{-i})$ and $\Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ denote the set of conditional probability systems on $(\Sigma_{-i}, \bar{\mathbb{H}}_i)$. Given a CPS $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$,

$$U_i^{\mu_i}(s_i, \theta_i, \mathcal{H}) = \int_{\Sigma_{-i}(\mathcal{H})} u_i(C(\zeta(s)), \theta) d\mu_i((s_{-i}, \theta_{-i})|\mathcal{H})$$

denotes agent i's expected utility if she plays strategy $s_i \in S_i(\mathcal{H})$, is of payoff type θ_i and holds beliefs $\mu_i(\cdot|\mathcal{H})$.

We call a strategy sequentially rational if it represents a sequentially rational plan of action. That is, we require a sequentially rational strategy to maximize expected utility at all information sets admitted by the strategy itself, but do not require optimality at information sets that cannot be reached if the agent follows the strategy. This is in line with the absence of belief-revision assumptions other than Bayesian updating,[4] standard in papers that define weak rationalizability or wPBE, and will play some role in our sufficiency mechanism.

Definition 2 *Strategy $s_i \in S_i$ is sequentially rational for payoff type $\theta_i \in \Theta_i$ of player i with respect to the beliefs $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ if for all $\mathcal{H} \in \mathbb{H}_i(s_i)$ and all $s_i' \in S_i(\mathcal{H})$*

$$U_i^{\mu_i}(s_i, \theta_i, \mathcal{H}) \geq U_i^{\mu_i}(s_i', \theta_i, \mathcal{H}). \tag{1}$$

We let $r_i : \Theta_i \times \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i}) \twoheadrightarrow S_i$ denote the correspondence that maps (θ_i, μ_i) to the set of strategies that are sequentially rational for payoff type θ_i with beliefs μ_i, and $\rho_i : \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i}) \twoheadrightarrow \Sigma_i$ denote the correspondence that maps μ_i to the set of strategy-payoff type pairs that includes (s_i, θ_i) if and only if s_i is sequentially rational for payoff type θ_i with beliefs μ_i.

2.4 Weak Rationalizability and WR-Implementation in Countable Mechanisms

A strategy is weakly rationalizable if it survives the iterative elimination of never-best sequential responses, where it is required that at the *initial* information set, each agent believes in the highest degree of her opponents' rationality. In static mechanisms, a strategy is weakly rationalizable if and only if is (belief-free) rationalizable as defined by BM.

[4] See e.g. Osborne and Rubinstein (1994) and Battigalli and Siniscalchi (2003) for a distinction between the terms "strategy" and "plan of action." In equilibrium concepts stronger than wPBE, such as sequential equilibrium, even after i deviates from her strategy, the other players are certain that i will follow her equilibrium strategy from now on. For such equilibrium concepts, one interpretation of the action specified by a strategy at an information set inconsistent with the strategy is not as the action planned by i, but rather as the other players' *belief* about i's behavior should i deviate (see e.g. Osborne and Rubinstein, 1994). In this context, requiring optimality of a strategy at *all* information sets implies that even after a deviation by i, the other players believe in equilibrium that i behaves optimally from now on. Weak rationalizability and wPBE do not impose such a belief-revision assumption. Thus, conceptually, the role of optimality at information sets inconsistent with a strategy is void.

Recall that Ord denotes the class of ordinal numbers.

Definition 3 *For all $i \in \mathcal{I}$, let $W_i^0 = \Sigma_i$ and $\Pi_i^0 = \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ and, for all ordinal numbers $\alpha \geq 1$, define by transfinite recursion the set W_i^α of weakly α-rationalizable pairs (s_i, θ_i) for player i by*

$$W_i^\alpha = \begin{cases} \rho_i(\Pi_i^{\alpha-1}) & \text{if } \alpha \text{ is a successor ordinal} \\ \bigcap_{\beta<\alpha} W_i^\beta & \text{if } \alpha \text{ is a limit ordinal} \end{cases}$$

and the set Π_i^α of weakly α-rationalizable beliefs for player i by

$$\Pi_i^\alpha = \left\{ \mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i}) : \mu_i(W_{-i}^\alpha | \{\emptyset\}) = 1 \right\}. \tag{2}$$

For convenience, let $Q_i^\alpha(\theta_i) = \{s_i \in S_i : (s_i, \theta_i) \in W_i^\alpha\}$ denote the set of weakly α-rationalizable strategies for $\theta_i \in \Theta_i$. Finally, let $W_i^\infty = \bigcap_{\alpha \in Ord} W_i^\alpha$ be the set of weakly rationalizable strategy-payoff type pairs for player i, $Q_i^\infty(\theta_i) = \{s_i \in S_i : (s_i, \theta_i) \in W_i^\infty\}$, and $\Pi_i^\infty = \bigcap_{\alpha \in Ord} \Pi_i^\alpha$ be the set of weakly rationalizable beliefs for player i.

For each agent i, Definition 3 iteratively eliminates pairs (s_i, θ_i) of never-best sequential responses and payoff types from the set $W_i^0 = \Sigma_i$ of all pairs of strategies and payoff types. If $(s_i, \theta_i) \in W_i^\alpha$ then (s_i, θ_i) survived α rounds of elimination. For $\alpha, \beta \in \text{Ord}$, $\alpha < \beta$ implies $W^\beta \subseteq W^\alpha$. The pair (s_i, θ_i) survives the first round of elimination if s_i is sequentially rational for θ_i with respect to some CPS from the set Π_i^0 of all CPSs. With each round of elimination the corresponding set of "permitted" CPSs weakly shrinks: for all $\alpha, \beta \in \text{Ord}$, $\alpha < \beta$ implies $\Pi^\beta \subseteq \Pi^\alpha$. In particular, in round α, Π_i^α consists of all CPSs μ_i which *initially* (that is, at $\{\emptyset\} \in \bar{\mathbb{H}}_i$) place probability one on W_{-i}^α. Thus each round increases the minimum level of rationality that i initially ascribes her opponents. Note that Π_i^α places no other restrictions on its elements. In particular, if $\mathcal{H}$ is a surprise given μ_i, then $\mu_i(\cdot|\mathcal{H})$ can be any probability distribution in $\Delta(\Sigma_{-i}(\mathcal{H}))$ without contradicting $\mu_i \in \Pi_i^\alpha$.

Battigalli (1999, 2003) and Battigalli and Siniscalchi (2007) define weak rationalizability as W^{ω_0}, where ω_0 denotes the first infinite ordinal. That is, they for all agents $i \in \mathcal{I}$ let $W_i^0 = \Sigma_i$ and $\Pi_i^0 = \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$, then for all *natural* numbers $\alpha \geq 1$ recursively define $W_i^\alpha = \rho_i(\Pi_i^{\alpha-1})$ and Π_i^α as in (2), and finally let $W_i^{\omega_0} = \bigcap_{\alpha=0}^\infty W_i^\alpha$ be the set of i's weakly rationalizable strategy-payoff type pairs. Definition 3 modifies their definition by requiring that one continues to iteratively eliminate never-best sequential best responses even after W^{ω_0} is reached.

Adding more rounds of elimination to Battigalli and Siniscalchi's definition does not affect our characterization of wr-implementable social choice functions.[5] However, the additional

[5]This will become clear after reading Section 3: Our proof by contradiction of the necessity of dr-monotonicity (Proposition 1) explicitly constructs a fixed point of the elimination procedure. The proof of the necessity of epIC (Proposition 2) would not change if we used Battigalli's and Siniscalchi's definition instead of Definition 3. Finally, as Footnote 14 will explain, the iterated elimination procedure of the mechanism

rounds of elimination are crucial for establishing an equivalence between robust wPBE- and wr-implementation. Briefly, the reason is that W^{ω_0} can fail to be a fixed-point of the elimination procedure of Definition 3, while W^∞ cannot. Correspondingly, for some agent i and payoff type θ_i, $Q_i^{\omega_0}(\theta_i)$ can contain a strategy that fails to be a wPBE strategy for θ_i in every type space, while $Q_i^\infty(\theta_i)$ cannot. We illustrate this point in Appendix A.

A social choice function $f : \Theta \to Y$ assigns a desired outcome to each payoff type profile. We pursue the full implementation of social choice functions. The key to fully weakly rationalizably implementing a social choice function f is to find a mechanism such that for every payoff type profile θ, *every* strategy profile that is weakly rationalizable for θ leads to $f(\theta)$.

Definition 4 *Mechanism Γ weakly rationalizably implements (wr-implements) social choice function f if*

(a) $C(\zeta(s)) = f(\theta)$ for all $(s, \theta) \in W^\infty$ and

(b) there exists a profile $(\mathcal{Q}_i(\theta_i))_{i\in\mathcal{I},\theta_i\in\Theta_i}$ of nonempty strategy sets $\mathcal{Q}_i(\theta_i) \subseteq Q_i^\infty(\theta_i)$ such that for all $i \in \mathcal{I}$, $\theta \in \Theta$ and $s_{-i} \in \mathcal{Q}_{-i}(\theta_{-i}) \overset{def}{=} \prod_{j\neq i} \mathcal{Q}_j(\theta_j)$, there exist $s_i \in \mathcal{Q}_i(\theta_i)$ and $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ such that $\mu_i((s_{-i}, \theta_{-i})|\{\emptyset\}) = 1$ and $s_i \in r_i(\theta_i, \mu_i)$.

Condition (a) is the standard requirement for full implementation described above. Condition (b) implies that every payoff type of every agent has some weakly rationalizable strategy, another standard requirement. In addition, Condition (b) demands the existence of a sequential best response to some weakly rationalizable CPSs that have a degenerate initial belief. In Section 4, we will see that Condition (b) corresponds to the requirement that an implementing mechanism has a wPBE for every type space. Thus Condition (b) ensures that wr-implementation has a foundation as robust wPBE-implementation.

3 Necessary and Sufficient Conditions for WR-Implementation

In full implementation results, preference reversal conditions often play an important role. In Subsection 3.1, we will introduce a novel preference reversal condition that is key to understanding wr-implementation. In Subsection 3.2 we will then use this condition to define dr-monotonicity, which Proposition 1 will establish as the main necessary condition for wr-implementation. In Subsection 3.3, Proposition 2 will reveal epIC as a second necessary condition. Finally, under a mild assumption, Subsections 3.4 and 3.5 will provide a converse to these necessary conditions. Specifically, Proposition 3 will show that under a conditional NTI condition, dr-monotonicity and epIC are also sufficient for wr-implementation.

that we employ to establish our sufficiency result (Proposition 3) converges in finitely many rounds.

In summary, with Propositions 1-3, this section will prove the following theorem. Note that Theorem 1 holds if we restrict attention to countable mechanisms but remains true if we admit general mechanisms.

Theorem 1 *Let f be a social choice function.*

(a) If f is wr-implementable then f is dr-monotone and epIC.

(b) If the conditional NTI property is satisfied and f is dr-monotone and epIC, then f is wr-implementable.

3.1 Robust Preference Reversals

Key to understanding dr-monotonicity, the central condition in Theorem 1, is to understand when what we will call a robust (θ_i', θ_i)-preference reversal exists.

Let $i \in \mathcal{I}$ be an agent and $\theta_i, \theta_i' \in \Theta_i$ be two payoff types of i. First, let us describe the following Condition (3) which, as we will show at the end of this subsection, guarantees the existence of a robust (θ_i', θ_i)-preference reversal. Imagine that i gets to pick exactly one outcome from a given set. Condition (3) is satisfied if and only if there is a subset of outcomes $Z \subseteq Y$ and a $y \in Z$ which θ_i would never choose from Z, independently of θ_i's belief ψ_i about θ_{-i}, but which θ_i' would choose from Z for some belief ψ_i' about θ_{-i}. To verify this claim, simply let Z equal the union of y and all different x from (3).

$$\exists \psi_i' \in \Delta(\Theta_{-i}), y \in Y \forall \psi_i \in \Delta(\Theta_{-i}) \exists x \in Y : \\ E_{\psi_i} u_i(x, \theta) > E_{\psi_i} u_i(y, \theta) \text{ and } E_{\psi_i'} u_i(x, \theta_i', \theta_{-i}) \leq E_{\psi_i'} u_i(y, \theta_i', \theta_{-i}). \quad (3)$$

Second, let us weaken Condition (3) to arrive at the preference reversal notion appropriate for our analysis. Recall that if $\psi_i' \in \Delta(\Theta_{-i})$, then $Y^{supp(\psi_i')}$ denotes the set of functions mapping $supp(\psi_i') = \{\theta_{-i} \in \Theta_{-i} : \psi_i'(\theta_{-i}) > 0\}$ to the set Y of outcomes. We say that there is a *robust (θ_i', θ_i)-preference reversal* if

$$\exists \psi_i' \in \Delta(\Theta_{-i}), y \in Y^{supp(\psi_i')} \forall \chi_i \in \Delta(supp(\psi_i') \times \Theta_{-i}) \exists x \in Y^{supp(\psi_i')} : \\ E_{\chi_i} u_i(x(\theta_{-i}'), \theta) > E_{\chi_i} u_i(y(\theta_{-i}'), \theta) \text{ and } E_{\psi_i'} u_i(x(\theta_{-i}), \theta_i', \theta_{-i}) \leq E_{\psi_i'} u_i(y(\theta_{-i}), \theta_i', \theta_{-i}).^6 \quad (4)$$

Condition (4) is in the same spirit as Condition (3), but recognizes that there are more elaborate ways to "separate" θ_i' from θ_i than to let i pick an element from a set Z. In particular, we can give choices to both i and $-i$. To understand Condition (4), imagine that i gets to choose the function y or one of the functions x and that $-i$ announces a payoff type profile

[6] Here, we let θ_{-i}' denote the first and θ_{-i} the second argument of χ_i, so that e.g. $E_{\chi_i} u_i(x(\theta_{-i}'), \theta) = \sum_{(\theta_{-i}', \theta_{-i}) \in supp(\psi_i') \times \Theta_{-i}} u_i(x(\theta_{-i}'), \theta) \chi_i(\theta_{-i}', \theta_{-i})$.

θ'_{-i}. Then, letting z denote i's choice, the outcome $z(\theta'_{-i})$ realizes. In this context, think of $\chi_i(\theta'_{-i}, \theta_{-i})$ as the probability that θ_i places on the event that $-i$'s actual payoff type profile is θ_{-i} but that $-i$ claims to be θ'_{-i}. Condition (4) then requires that θ_i would never choose y, independently of θ_i's belief χ_i, but that θ'_i would choose y if she believes that θ_{-i} is distributed according to ψ'_i and that $-i$ tells the truth.

As already mentioned, Condition (3) is sufficient for the existence of a robust (θ'_i, θ_i)-preference reversal: Condition (3) is satisfied if in Condition (4), we can choose y and all x to be constant functions. In this special case i's belief about $-i$'s announced type profile is irrelevant, and the beliefs χ_i about $-i$'s announced and actual payoff type profiles from (4) can be replaced in (3) by the beliefs ψ_i about $-i$'s actual payoff type profile only.

3.2 Dynamic Robust Monotonicity and its Necessity

Let f be the social choice function under consideration. In order to define our first necessary condition for wr-implementation, we recall the notion of a deception. A *deception* is a profile $\beta = (\beta_1, \ldots, \beta_I)$, where $\beta_i : \Theta_i \to 2^{\Theta_i}$ satisfies $\theta_i \in \beta_i(\theta_i)$ for all $i \in \mathcal{I}$ and all $\theta_i \in \Theta_i$. We can interpret a deception β in terms of the direct mechanism associated with f.[7] In this context, $\beta_i(\theta_i)$ is the set of i's strategies that θ_i can play under β. A deception β is *acceptable* if for all $\theta, \theta' \in \Theta$, $\theta' \in \prod_{j \in \mathcal{I}} \beta_j(\theta_j)$ implies $f(\theta') = f(\theta)$, and *unacceptable* otherwise. For each $\vartheta_{-i} \in \Theta_{-i}$, $\beta_{-i}^{-1}(\vartheta_{-i}) = \{\theta_{-i} \in \Theta_{-i} : \vartheta_{-i} \in \beta_{-i}(\theta_{-i})\}$ is the set of payoff type profiles that can announce ϑ_{-i} under β, where $\beta_{-i}(\theta_{-i}) \overset{\text{def}}{=} \prod_{j \neq i} \beta_j(\theta_j)$.

More broadly, in a mechanism that may be direct or indirect, we can think of $\beta_i(\theta_i)$ as the set of payoff types that θ_i is permitted to imitate under β. Intuitively, if we want to achieve wr-implementation, then in the implementing mechanism, every unacceptable deception may not form a fixed point of the iterated elimination procedure defining weak rationalizability. In order for it not to form a fixed point, an unacceptable deception β must permit some payoff type θ_i of some agent i a lie θ'_i that is not worth abiding by if the payoff type initially believes that the other agents follow β. In fact, β must be d-refutable in the sense defined below, where, like BM, we let

$$Y_i(\theta_{-i}) = \left\{ y \in Y : u_i(y, (\theta''_i, \theta_{-i})) \leq u_i(f(\theta''_i, \theta_{-i}), (\theta''_i, \theta_{-i})) \text{ for all } \theta''_i \in \Theta_i \right\}$$

be the "reward set" for agent i (with respect to θ_{-i}).

Definition 5 *A deception β is dynamically refutable (d-refutable) if there exist $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta'_i \in \beta_i(\theta_i)$ such that 1) there is a robust (θ'_i, θ_i)-preference reversal or 2) for all $\theta'_{-i} \in \Theta_{-i}$*

[7] The direct mechanism associated with the social choice function f is the static mechanism Γ with $S_i = \Theta_i$ for all $i \in \mathcal{I}$ and $C(\theta) = f(\theta)$ for all $\theta \in \Theta$.

and all $\psi_i \in \Delta(\Theta_{-i})$ *with* $\psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1$,

$$\exists x \in Y_i(\theta'_{-i}) : E_{\psi_i} u_i(x, \theta) > E_{\psi_i} u_i(f(\theta'), \theta). \tag{5}$$

Intuitively, d-refutability captures when a deception β can be "handled" via the use of some dynamic mechanism. In a condition termed refutability, BM already captured when a deception can be handled by some static mechanism. Our formulation of d-refutability permits a straightforward comparison between these two notions: a deception β is refutable according to BM if there exist $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta'_i \in \beta_i(\theta_i)$ such that 2) holds. Hence a mechanism designer who considers not only static but also dynamic mechanisms gains that she can now also handle a deception β if there exist $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta'_i \in \beta_i(\theta_i)$ such that 1) holds. This gain applies even if nothing (beyond Bayesian updating) is known about the agents' belief revision processes.

A high-level intuition for d-refutability is the following. (The proof of Theorem 1 will account for all details.) In order to handle a deception β by disincentivizing θ_i's lie θ'_i, a mechanism that implements a social choice function needs to have two properties.

- The mechanism must permit θ_i a deviation that leads to a more favorable expected outcome than that obtained by imitating θ'_i.
- As Part (b) of Definition 4 roughly means that truth-telling must form a fixed point in every implementing mechanism, the existence of the more favorable expected outcome must not upset the truth-telling fixed point.

A dynamic mechanism offers two ways to disincentivize θ_i's lie $\theta'_i \in \beta_i(\theta_i)$. First, it may disincentivize θ'_i at a surprise information set, and second, at the initial or a similar information set.

At a surprise information set, since we impose no belief-revision assumptions, we do not know anything about the agent's beliefs. Consequently, a dynamic mechanism "separates" θ'_i from θ_i by their behavior at a surprise information set exactly if there is a robust (θ'_i, θ_i)-preference reversal. In particular, the two bullet points above correspond to the two inequalities in (4): There is a choice y that, no matter her belief about her opponents, θ_i will never make in the "subgame" starting at the surprise information set (first bullet point above and strict inequality in (4)), but that θ'_i is happy to make if she believes in her opponents telling the truth (second bullet point above and weak inequality in (4)).

If no robust (θ'_i, θ_i)-preference reversal exists, then the mechanism must disincentivize θ_i's lie $\theta'_i \in \beta_i(\theta_i)$ at the initial (or a similar) information set. Let us first consider the case of a static mechanism, already examined by BM. The only option to achieve in a static mechanism that θ_i does not imitate θ'_i is to give player i, for every belief of θ_i about her opponents that

conforms to β, an action m that is more attractive to θ_i than θ_i''s weakly rationalizable actions. In particular, if θ_i expects her opponents to play θ'_{-i} and believes that their actual payoff type profiles are distributed according to ψ_i, then for some such additional action m, the action profile (m, θ'_{-i}) must lead to some outcome x that satisfies the strict inequality in (5) (first bullet point above). At the same time, the additional action m must not upset the truth-telling fixed point (second bullet point above): if i expects θ'_{-i} to tell the truth, then, no matter her payoff type θ_i'', agent i must not prefer playing m and receiving x over telling the truth. This is captured by the condition that x must be in the reward set $Y_i(\theta'_{-i})$. Second, while dynamic mechanisms offer more options than adding an extra action to i's initial information set (e.g., our sufficiency mechanism from Proposition 3 will augment the direct mechanism in a dynamic fashion), the static "whistleblower" condition 2) continues to be necessary even with dynamic mechanisms in case 1) does not hold. The intuition is that 2) now captures the possibility to disincentivize θ_i's lie θ_i' by, for each θ'_{-i}, adding actions at information sets that i initially expects with strictly positive probability if she believes the other players imitate θ'_{-i}.

Observe two differences between 1) and 2) that arise from 1) capturing the possibility to disincentivize θ_i's lie θ_i' at a surprise information set, and 2) capturing the same at the initial or a similar information set. First, corresponding to the lack of knowledge about the agent's beliefs after a surprise, the preference reversal described by 1) is robust, that is, holds for all possible beliefs of θ_i. In contrast, 2) requires a reversal only for beliefs that conform to the deception β. Second, however, corresponding to a surprise being unexpected, 1) makes no restriction on the (expected) outcomes x and y with respect to which there is a preference reversal. In contrast, 2) requires these outcomes to be in the reward set. Therefore, 1) and 2) are complementary in that they separate θ_i' from θ_i in different circumstances.

Definition 6 *Social choice function f is dynamically robustly monotone (dr-monotone) if every unacceptable deception is d-refutable.*

Proposition 1 *If social choice function f is wr-implementable then f is dr-monotone.*

BM, Theorem 1, Corollary 1 show that a condition related to dr-monotonicity, robust monotonicity, is necessary for wr-implementation in static mechanisms.[8] A social choice function is robustly monotone if every unacceptable deception is refutable. Robust monotonicity is more demanding than dr-monotonicity.

We prove Proposition 1 in Appendix B. There, we suppose by contradiction that some Γ wr-implements f but that there is a deception β which is unacceptable but not d-refutable. We then construct a fixed point corresponding to β of the iterated elimination procedure that defines weak rationalizability. Our construction exploits that if β permits a payoff type θ_i

[8]More precisely, they derive that *strict* robust monotonicity (and hence robust monotonicity) is necessary for their notion of rationalizable implementation, and hence for robust implementation in static mechanisms.

to mimic another payoff type θ_i', then the lack of d-refutability of β implies that there is no robust (θ_i', θ_i)-preference reversal. Specifically, the absence of a robust (θ_i', θ_i)-preference reversal helps us construct the non-initial beliefs of a CPS that justifies the lie θ_i' for θ_i. Since β is unacceptable, the fixed point implies that Γ produces an undesired outcome, contradicting that Γ wr-implements f.

3.3 Ex-post Incentive Compatibility and its Necessity

Bergemann and Morris (2005) show that ex-post incentive compatibility (epIC) is a necessary condition of partial robust implementation in static mechanisms.

Definition 7 *Social choice function f is epIC if for all $i \in \mathcal{I}$, $\theta_i, \theta_i' \in \Theta_i$ and all $\theta_{-i} \in \Theta_{-i}$*

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

Partial robust implementation only requires that for each type space, there *exists* a wPBE that delivers the social choice. If instead one pursues the more demanding notion of full robust implementation, that is, if one also insists that for each type space, *all* wPBE deliver the social choice, a stronger version of epIC called semi-strict epIC emerges as a necessary condition for robust implementation in static mechanisms (BM). For (full) wr-implementation in dynamic mechanisms, on the other hand, the necessary incentive compatibility condition remains epIC (whether we restrict attention to countable dynamic mechanisms or not).

Proposition 2 *If social choice function f is wr-implementable then f is epIC.*

In the static case, the necessary condition of robust monotonicity implies semi-strict epIC (BM, Lemma 1, Theorem 1). The following example shows that despite that, an analogous relation does not hold between dr-monotonicity and epIC. Hence we need to prove Proposition 2 directly, and cannot deduce it from Proposition 1. We do so in Appendix B.

Example 3.1 There is one agent with two payoff types, $\Theta_1 = \{\theta_1, \theta_1'\}$. (We could easily add a "dummy" second agent with $\Theta_2 = \{\theta_2\}$ in order to maintain our assumption of at least two agents from Subsection 2.1.) There are two pure outcomes, $X = \{x, x'\}$. Payoff type θ_1 strictly prefers x over x', while θ_1' has the opposite strict preferences:

$$u_1(x, \theta_1) > u_1(x', \theta_1) \quad \text{and} \quad u_1(x, \theta_1') < u_1(x', \theta_1').$$

Therefore, both a robust (θ_1', θ_1)- and a robust (θ_1, θ_1')-preference reversal exist. Consider the social choice function $f : \Theta_1 \to Y$ given by $f(\theta_1) = x'$ and $f(\theta_1') = x$. This social choice function is not epIC, but nonetheless dr-monotone since by the existence of the robust preference reversals, every unacceptable deception is d-refutable.

3.4 Sufficient Conditions

For the purpose of deriving sufficient conditions for wr-implementation, we assume that players' preferences satisfy the following mild condition, taken from BM.

Definition 8 (Conditional NTI.) *Let f be a social choice function. The conditional no total indifference (NTI) property is met if for all $i \in \mathcal{I}$, $\theta_i \in \Theta_i$, $\psi_i \in \Delta(\Theta_{-i})$ and $\theta'_{-i} \in \Theta_{-i}$, there exist $y, y' \in Y_i(\theta'_{-i})$ such that*

$$E_{\psi_i} u_i(y, \theta) > E_{\psi_i} u_i(y', \theta).$$

Strictly speaking, the conditional NTI is a property of the social choice function f under consideration. This is because the reward sets $Y_i(\theta'_{-i})$ depend on f. However, in many cases of interest (e.g. quasilinear environments) the conditional NTI is satisfied by all social choice functions and in this sense, can be thought of as a property of the environment.

Proposition 3 *If social choice function f is dr-monotone and epIC and the conditional NTI property is satisfied, then f is wr-implementable.*

We prove Proposition 3 in the following subsection using an infinite mechanism. Our mechanism has two debatable properties. First, it exploits that faced with an infinite number of choices, an agent may not have a sequential best response if she holds certain beliefs about her opponents. This property is familiar from the literature, including from the static results by BM. Second and possibly less common, our mechanism leverages the plans of action notion of sequential rationality embedded in weak rationalizability (see Definition 2 and the paragraph preceding it). Concretely, it exploits that in a dynamic mechanism, an agent may avoid strategies that admit information sets at which the agent has no best response. In Müller (2017b), we show that under a mild strengthening of our sufficient conditions, this second property can be dispensed with at the cost of complicating the early stages of the implementing mechanism.

The use of non-well-behaved infinite mechanisms has been criticized by Jackson (1992) and others but allows us to derive a clean result, while a characterization of wr-implementation by well-behaved mechanisms remains an open and challenging question. Since non-well-behaved infinite mechanism are common in papers on full implementation, our choice also allows an easy comparison of our sufficiency result to the existing literature.

3.5 Proof of Proposition 3

Proof. First, some preliminaries.

Relative Interior Lotteries of the Reward Sets. For every $i \in \mathcal{I}$ and every $\theta_{-i} \in \Theta_{-i}$, the reward set $Y_i(\theta_{-i})$ is the intersection of $Y = \{y \in \mathbb{R}^{\#X} : y \geq 0, \sum y_n = 1\}$ with the half-spaces $\{(y_x)_{x\in X} \in \mathbb{R}^{\#X} : \sum_{x\in X} u_i(x, \theta_i', \theta_{-i}) y_x \leq u_i(f(\theta_i', \theta_{-i}), \theta_i', \theta_{-i})\}$, indexed by $\theta_i' \in \Theta_i$. As such, $Y_i(\theta_{-i})$ is convex and has finitely many extreme points $y_{1,\theta_{-i}}, y_{2,\theta_{-i}}, \ldots, y_{m,\theta_{-i}}$. By the conditional NTI property, $Y_i(\theta_{-i})$ is nonempty. Therefore $m \geq 1$, and it is well-defined to let $\bar{y}_{\theta_{-i}}$ be the convex combination that puts weight $\frac{1}{m}$ on each extreme point. The "relative interior" lotteries $\bar{y}_{\theta_{-i}} \in Y_i(\theta_{-i})$, indexed by $i \in \mathcal{I}$ and $\theta_{-i} \in \Theta_{-i}$, satisfy the following two properties.

- (compare with BM, p. 270) There is a $(Y_i^F(\theta'_{-i}))_{i\in\mathcal{I},\theta'_{-i}\in\Theta_{-i}}$ such that for each $i \in \mathcal{I}$ and $\theta'_{-i} \in \Theta_{-i}$, $Y_i^F(\theta'_{-i})$ is a finite subset of the reward set $Y_i(\theta'_{-i})$, and

$$\forall i \in \mathcal{I}, \theta_i \in \Theta_i, \theta'_{-i} \in \Theta_{-i}, \psi_i \in \Delta(\Theta_{-i})\ \exists y \in Y_i^F(\theta'_{-i}) : E_{\psi_i} u_i(y, \theta) > E_{\psi_i} u_i(\bar{y}_{\theta'_{-i}}, \theta). \tag{6}$$

  Proof: Let $i \in \mathcal{I}$ and $\theta'_{-i} \in \Theta_{-i}$. By the conditional NTI property, for every $\theta_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i})$, there are $y = \sum \alpha_k y_{k,\theta'_{-i}} \in Y_i(\theta'_{-i})$ and $y' = \sum \alpha'_k y_{k,\theta'_{-i}} \in Y_i(\theta'_{-i})$ such that $E_{\psi_i} u_i(y, \theta) > E_{\psi_i} u_i(y', \theta)$. For $\eta > 0$ small enough, $\bar{y}_{\theta'_{-i}} + \eta(y - y') = \sum \left(\frac{1}{m} + \eta(\alpha_k - \alpha'_k)\right) y_{k,\theta'_{-i}}$ is in $Y_i(\theta'_{-i})$. Moreover, $E_{\psi_i} u_i(\bar{y}_{\theta'_{-i}} + \eta(y - y'), \theta) > E_{\psi_i} u_i(\bar{y}_{\theta'_{-i}}, \theta)$. Thus $\forall \theta_i, \psi_i\ \exists y(\theta_i, \psi_i) \in Y_i(\theta'_{-i}) : E_{\psi_i} u_i(y(\theta_i, \psi_i), \theta) > E_{\psi_i} u_i(\bar{y}_{\theta'_{-i}}, \theta)$. Endow $\Delta(\Theta_{-i})$ with the topology of weak convergence and recall that Θ_j is finite for all $j \in \mathcal{I}$. Then for each θ_i and ψ_i, there exists an open set $O_{\psi_i} \subseteq \Delta(\Theta_{-i})$ such that $\psi_i \in O_{\psi_i}$ and such that $E_{\psi_i'} u_i(y(\theta_i, \psi_i), \theta) > E_{\psi_i'} u_i(\bar{y}_{\theta'_{-i}}, \theta)$ for all $\psi_i' \in O_{\psi_i}$. Moreover, $\Delta(\Theta_{-i})$ is compact, and there exists a finite subcover $(O_{\psi_i^k})_{k=1}^m$ of the open cover (O_{ψ_i}) of $\Delta(\Theta_{-i})$. Now let $Y_i^F(\theta'_{-i}) = \{y(\theta_i, \psi_i^k) : \theta_i \in \Theta_i, k \in \{1, \ldots, m\}\}$. If we define $Y_i^F(\theta'_{-i})$ in this way for all i and θ'_{-i}, then $(Y_i^F(\theta'_{-i}))_{i\in\mathcal{I},\theta'_{-i}\in\Theta_{-i}}$ satisfies (6).

- There exists $\varepsilon \in (0, 1)$ such that for all $i \in \mathcal{I}$, $\theta \in \Theta$ and $y \in Y$

$$u_i\big((1-\varepsilon)\bar{y}_{\theta_{-i}} + \varepsilon y, \theta\big) < u_i(f(\theta), \theta). \tag{7}$$

  Proof: By the conditional NTI property, for every i, θ, there are $y = \sum \alpha_k y_{k,\theta_{-i}} \in Y_i(\theta_{-i})$ and $y' = \sum \alpha'_k y_{k,\theta_{-i}} \in Y_i(\theta_{-i})$ such that $u_i(y, \theta) > u_i(y', \theta)$. By definition of $Y_i(\theta_{-i})$, $u_i(y_{k,\theta_{-i}}, \theta) \leq u_i(f(\theta), \theta)$ for all $k = 1, \ldots, m$. Suppose that $u_i(y_{k,\theta_{-i}}, \theta) = u_i(f(\theta), \theta)$ for all k, then

$$u_i(y, \theta) = \sum \alpha_k u_i(y_{k,\theta_{-i}}, \theta) = u_i(f(\theta), \theta) = \sum \alpha'_k u_i(y_{k,\theta_{-i}}, \theta) = u_i(y', \theta).$$

  Contradiction, hence $u_i(y_{\bar{k},\theta_{-i}}, \theta) < u_i(f(\theta), \theta)$ for some $\bar{k}$. Therefore, for all $i \in \mathcal{I}$ and $\theta \in \Theta$, $u_i(\bar{y}_{\theta_{-i}}, \theta) < u_i(f(\theta), \theta)$. Since X is finite, for all $i \in \mathcal{I}$ and $\theta \in \Theta$, there exists

$\varepsilon_i(\theta) \in (0,1)$ such that for all $y \in Y$, $u_i\left((1-\varepsilon_i(\theta))\overline{y}_{\theta_{-i}} + \varepsilon_i(\theta)y, \theta\right) < u_i(f(\theta),\theta)$. Since Θ and $\mathcal{I}$ are finite, (7) follows.

D-refutable Deceptions. By an argument analogous to that made in the proof of (6), the compactness in the topology of weak convergence of $\{\psi_i \in \Delta(\Theta_{-i}) : \psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1\}$ for every deception β and every $\theta'_{-i} \in \Theta_{-i}$ implies the following equivalence: A deception β is d-refutable if and only if there exist $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta'_i \in \beta_i(\theta_i)$ such that 1) there is a robust (θ'_i, θ_i)-preference reversal or 2) for all $\theta'_{-i} \in \Theta_{-i}$ there is a finite subset $Y_i^F(\theta'_{-i})$ of the reward set $Y_i(\theta'_{-i})$ such that for all $\psi_i \in \Delta(\Theta_{-i})$ with $\psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1$,

$$\exists x \in Y_i^F(\theta'_{-i}) : E_{\psi_i} u_i(x,\theta) > E_{\psi_i} u_i(f(\theta'),\theta). \tag{8}$$

For each d-refutable deception β let $i = i(\beta) \in \mathcal{I}$, $\theta_i = \theta_i(\beta) \in \Theta_i$ and $\theta'_i = \theta'_i(\beta) \in \beta_i(\theta_i)$ be such that 1) or 2) are true. For every d-refutable β for which 2) holds, every $\theta'_{-i} \in \Theta_{-i}$ and every $\psi_i \in \Delta(\Theta_{-i})$ such that $\psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1$ let $x^{(\beta,\theta'_{-i},\psi_i)} \in Y_i^F(\theta'_{-i})$ be such that (8) holds.

Finite Subsets of Reward Sets. For each i and a θ'_{-i}, we defined a set $Y_i^F(\theta'_{-i})$ that satisfies (6). Moreover, for each d-refutable β, if $i = i(\beta)$ and 2) holds for θ'_{-i}, the previous paragraph introduced another set $Y_i^F(\theta'_{-i})$. For the remainder of the proof, $Y_i^F(\theta'_{-i})$ will denote the union of all these sets, which is still a finite set and satisfies both (8) for all pertinent β and (6).

Robust Preference Reversals. For every $(\theta'_i, \theta_i) \in \Theta_i^2$ for which there is a robust (θ'_i, θ_i)-preference reversal, let $\psi'_i = \psi'_i(\theta'_i, \theta_i) \in \Delta(\Theta_{-i})$, $y(\theta'_i, \theta_i) \in Y^{supp(\psi'_i)}$, $p = p(\theta'_i, \theta_i) \in \mathbb{N}$, $x^1(\theta'_i, \theta_i), \ldots, x^p(\theta'_i, \theta_i) \in Y^{supp(\psi'_i)}$ be such that for all $\chi_i \in \Delta(supp(\psi'_i) \times \Theta_{-i})$ there exists a $q \in \{1, \ldots, p\}$ such that

$$\begin{aligned} &E_{\chi_i} u_i(x^q(\theta'_i,\theta_i)(\theta'_{-i}),\theta) > E_{\chi_i} u_i(y(\theta'_i,\theta_i)(\theta'_{-i}),\theta) \\ &\qquad \text{and } E_{\psi'_i} u_i(x^q(\theta'_i,\theta_i)(\theta_{-i}),\theta'_i,\theta_{-i}) \leq E_{\psi'_i} u_i(y(\theta'_i,\theta_i)(\theta_{-i}),\theta'_i,\theta_{-i}). \end{aligned} \tag{9}$$

Similar to above, such p and $x^1(\theta'_i, \theta_i), \ldots, x^p(\theta'_i, \theta_i)$ exist because $\Delta(supp(\psi'_i) \times \Theta_{-i})$ is compact.

Given these preliminaries, we now construct a mechanism $\Gamma = \langle A, H, (\mathbb{H}_i)_{i\in\mathcal{I}}, P, C\rangle$ which we might describe as a "dynamically augmented" direct mechanism. For ease of reference, we divide Γ into four stages. Figures 1 and 2 present a scheme equivalent to Γ.

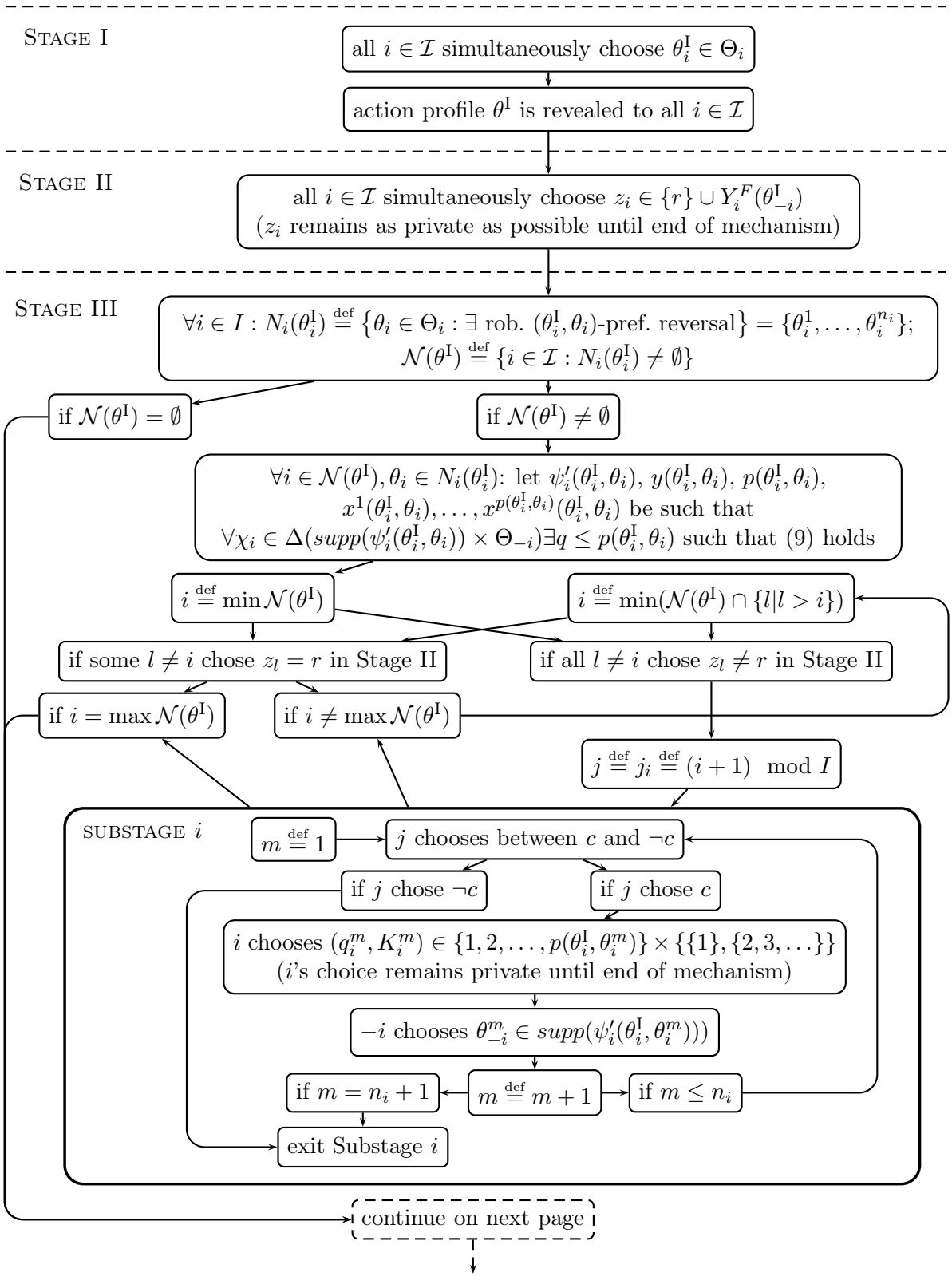

Figure 1: Stages I, II and III of the mechanism of Proposition 3

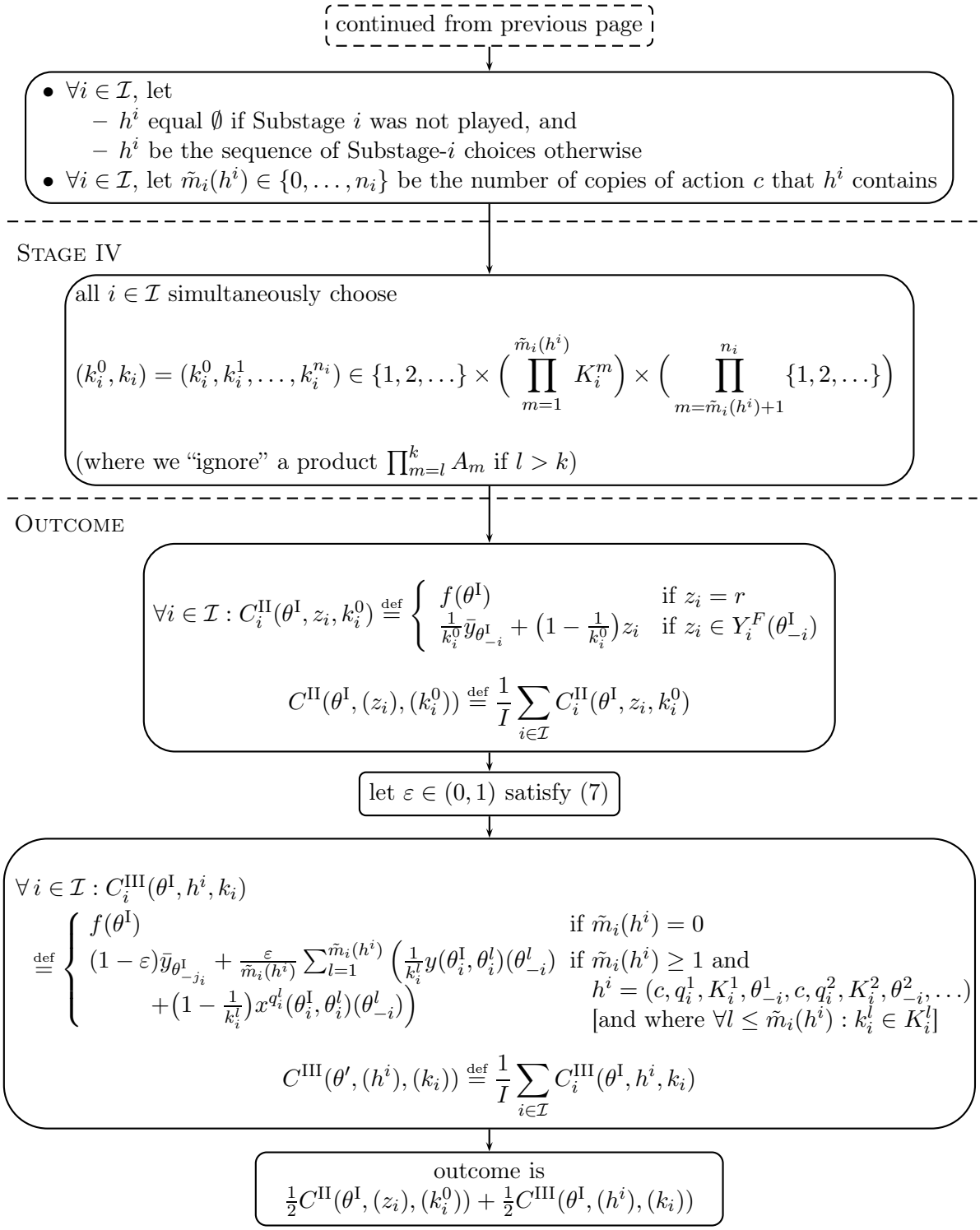

Figure 2: Stage IV and outcome function of the mechanism of Proposition 3

Stage I. In the first stage of Γ, each agent i announces one of her payoff types $\theta_i^{\mathrm{I}} \in \Theta_i$, strategically simultaneously with all other the agents.[9] The set of first-stage "subhistories" is $H^{\mathrm{I}} = \{h : h \preceq \theta \text{ for some } \theta \in \Theta\}$. In the following paragraphs, we will also describe subhistories for the Stages II, III and IV; terminal histories in Γ will be concatenations of "terminal subhistories." Every agent i has exactly one first-stage information set, namely $\mathcal{H}_i^{\mathrm{I}} = \{(\theta_1, \ldots, \theta_{i-1}) \in H^{\mathrm{I}}\} \in \mathbb{H}_i$.[10] Accordingly, if $h = (\theta_1, \ldots, \theta_{i-1}) \in H^{\mathrm{I}}$ then $P(h) = i$. (As defining the player function P for Stage II-IV histories is similarly trivial, we will omit P's definition from now on.) All players observe the first-stage action profile θ^{I} before the second stage commences.

Stage II. Given the first-stage action profile $\theta^{\mathrm{I}} \in \Theta$, every agent i chooses a second-stage message $z_i \in \{r\} \cup Y_i^F(\theta_{-i}^{\mathrm{I}})$, strategically simultaneously with all other agents. Here, r stands for "rational" (Step 1 of this proof will imply that r is the only rational choice for i among the actions z_i). Thus, we can write terminal second-stage subhistories that follow the terminal first-stage subhistory θ^{I} as

$$(z_i)_{i \in \mathcal{I}} \in \prod_{i \in \mathcal{I}} \left(\{r\} \cup Y_i^F(\theta_{-i}^{\mathrm{I}}) \right).$$

Every player's second-stage action remains as private as possible throughout the mechanism, as page 24 will explain further. There, we will also define the agents' second-stage information sets.

Stage III. For all $i \in \mathcal{I}$, let $N_i(\theta_i^{\mathrm{I}})$ be the set of $\theta_i \in \Theta_i$ for which a robust $(\theta_i^{\mathrm{I}}, \theta_i)$-preference reversal exists (recall that we know that $\theta_i \in N_i(\theta_i^{\mathrm{I}})$ if there is a d-refutable β such that $i = i(\beta)$, $\theta_i = \theta_i(\beta)$ and $\theta_i^{\mathrm{I}} = \theta_i'(\beta)$ for which 2) is false for (i, θ_i, θ_i')). Number the elements of $N_i(\theta_i^{\mathrm{I}})$ in a random order and write $N_i(\theta_i^{\mathrm{I}}) = \{\theta_i^1, \ldots, \theta_i^{n_i}\}$, with the convention that $n_i = 0$ if $N_i(\theta_i^{\mathrm{I}}) = \emptyset$. Let $\mathcal{N}(\theta^{\mathrm{I}}) = \{i \in \mathcal{I} : N_i(\theta_i^{\mathrm{I}}) \neq \emptyset\}$.

Suppose that $\mathcal{N}(\theta^{\mathrm{I}}) \neq \emptyset$. Then Stage III has up to $\#\mathcal{N}(\theta^{\mathrm{I}})$ "substages," which we label by the elements of $\mathcal{N}(\theta^{\mathrm{I}})$. The first of these (potential) substages is substage $i = \min \mathcal{N}(\theta^{\mathrm{I}})$. It is played if all $j \neq i$ chose some $z_j \in Y_j^F(\theta_{-j}^{\mathrm{I}})$ in Stage II, and skipped if some $j \neq i$ chose r in Stage II. In either case, substage i is followed by (potential) substage $i' = \min \mathcal{N}(\theta^{\mathrm{I}}) \backslash \{i\}$. Substage i' is played if all $j \neq i'$ chose some $z_j \in Y_j^F(\theta_{-j}^{\mathrm{I}})$ in Stage II, and skipped otherwise. In either case, substage i' is followed by (potential) substage $i'' = \min \mathcal{N}(\theta^{\mathrm{I}}) \backslash \{i, i'\}$. And so on.

Let us now describe substage $i \in \mathcal{N}(\theta^{\mathrm{I}})$, assuming that it is played and not skipped. Substage i starts with $j_i = (i+1) \mod I$ choosing between two actions c (for "continue")

[9] This assumes that every agent has at least two payoff types. For each $i \in \mathcal{I}$, if Θ_i is a singleton, then to ensure that Γ satisfies the no trivial decision condition (see Definition 12 in Subsection 4.2) of a mechanism, i does not take part in Stage I and obvious notational changes ensue.

[10] In this proof, we use concise notation for information sets. E.g., since in the context of defining $\mathcal{H}_i^{\mathrm{I}}$ the tuple $(\theta_1, \ldots, \theta_{i-1})$ does not denote a specific profile of payoff types, $\{(\theta_1, \ldots, \theta_{i-1}) \in H^{\mathrm{I}}\}$ denotes $\{(\theta_1, \ldots, \theta_{i-1}) \in H^{\mathrm{I}} : (\theta_1, \ldots, \theta_{i-1}) \in \prod_{j<i} \Theta_j\}$.

and $\neg c$ (for "not continue").[11] Action $\neg c$ ends the substage, while if c is played then player i chooses a $(q_i^1, K_i^1) \in \{1, 2, \ldots, p(\theta_i^{\mathrm{I}}, \theta_i^1)\} \times \{\{1\}, \{2, 3, \ldots\}\}$. To anticipate, in Stage IV, i will pick a $k_i^1 \in \{1, 2, 3, \ldots\}$ such that $k_i^1 \in K_i^1$. Thus, in Stage III, i essentially decides between $k_i^1 = 1$ and $k_i^1 \geq 2$, and in the latter case she will refine her choice in Stage IV. Following i's choice, "$-i$ chooses a $\theta_{-i}^1 \in supp(\psi_i'(\theta_i^{\mathrm{I}}, \theta_i^1))$:" first, $l_1 = \min \mathcal{I} \backslash \{i\}$ chooses a $\theta_{l_1}^1 \in \Theta_{l_1}$ such that $\theta_{l_1}^1 \in supp(marg_{\Theta_{l_1}} \psi_i'(\theta_i^{\mathrm{I}}, \theta_i^1))$, then $l_2 = \min \mathcal{I} \backslash \{i, l_1\}$ chooses a $\theta_{l_2}^1 \in \Theta_{l_2}$ such that $(\theta_{l_1}^1, \theta_{l_2}^1) \in supp(marg_{\Theta_{l_1} \times \Theta_{l_2}} \psi_i'(\theta_i^{\mathrm{I}}, \theta_i^1))$, and so on, until all $l \neq i$ have chosen.[12] After these choices, if $\#N_i(\theta_i') = 1$, the substage ends. If $\#N_i(\theta_i') > 1$, then a second round of choices follows: player j_i again chooses between two actions c and $\neg c$. Action $\neg c$ ends the substage, while if c is played then i chooses a $(q_i^2, K_i^2) \in \{1, 2, \ldots, p(\theta_i^{\mathrm{I}}, \theta_i^2)\} \times \{\{1\}, \{2, 3, \ldots\}\}$, followed by "$-i$ choosing a $\theta_{-i}^2 \in supp(\psi_i'(\theta_i^{\mathrm{I}}, \theta_i^2))$." If $\#N_i(\theta_i') = 2$ the substage ends while if $\#N_i(\theta_i') > 2$, a further round of choices ensues. And so on. If j_i does not end the substage by playing $\neg c$ before, the substage ends after i chooses $(q_i^{n_i}, K_i^{n_i})$ and "$-i$ chooses $\theta_{-i}^{n_i}$."

Suppose that the terminal first- and second-stage subhistories are θ^{I} and (z_i), respectively. Let us write terminal third-stage subhistories histories as $h^{\mathrm{III}} = (h^i)$, where we subdivide h^{III} into terminal Substage i subhistories h^i, and let (h^i) denote the concatenation of these substage subhistories. Here, let $h^i = \emptyset$ if there is no Substage i given θ^{I} because $i \notin \mathcal{N}(\theta^{\mathrm{I}})$ or if $i \in \mathcal{N}(\theta^{\mathrm{I}})$ but Substage i was skipped because some $j \neq i$ played $z_j = r$ in Stage II. Let $\tilde{m}_i(h^i) \in \{0, \ldots, n_i\}$ denote the number of copies of action c that h^i contains. That is, $\tilde{m}_i(h^i) = 0$ if $h^i \in \{\emptyset, (\neg c)\}$, $\tilde{m}_i(h^i) = \tilde{m}$ if

$$h^i \in \prod_{m \leq \tilde{m}} \Big(\{c\} \times \{1, \ldots, p(\theta_i^{\mathrm{I}}, \theta_i^m)\} \times \{\{1\}, \{2, 3, \ldots\}\} \times supp(\psi_i'(\theta_i^{\mathrm{I}}, \theta_i^m)) \Big) \times \{\neg c\},$$

and $\tilde{m}_i(h^i) = n_i$ if h^i is a $(4n_i)$-tuple $(c, q_i^1, K_i^1, \theta_{-i}^1, \ldots, c, q_i^{n_i}, K_i^{n_i}, \theta_{-i}^{n_i})$.

Stage IV. As a final choice, given the history of play $(\theta^{\mathrm{I}}, (z_i), (h^i))$ from the previous three stages, each agent i chooses a pair (k_i^0, k_i) consisting of a number k_i^0 and a tuple k_i, strategically simultaneously with all other agents. More precisely, agent i chooses

$$(k_i^0, k_i) = (k_i^0, k_i^1, \ldots, k_i^{n_i}) \in \{1, 2, \ldots\} \times \Big(\prod_{m=1}^{\tilde{m}_i(h^i)} K_i^m \Big) \times \Big(\prod_{m=\tilde{m}_i(h^i)+1}^{n_i} \{1, 2, \ldots\} \Big),$$

where $n_i = 0$ (which implies $h^i = \emptyset$ and thus $\tilde{m}_i(h^i) = 0$) is understood to imply that i only chooses k_i^0, $\tilde{m}_i(h^i) = 0$ and $n_i > 0$ that $(k_i^0, k_i) \in \{1, 2, \ldots\}^{n_i+1}$, and $\tilde{m}_i(h^i) = n_i > 0$ that $(k_i^0, k_i) \in \{1, 2, \ldots\} \times \Big(\prod_{m=1}^{n_i} K_i^m \Big)$.

[11] It works equally well to let j_i be any other player different from i.

[12] To ensure that Γ satisfies the no trivial decision condition, for every $l \neq i$, if there is only one $\theta_l^1 \in \Theta_l$ such that $(\theta_j^1)_{j \leq l, j \neq i} \in supp(marg_{(\Theta_j)_{j \leq l, j \neq i}} \psi_i'(\theta_i', \theta_i^1))$, then we skip l. That is, immediately after $l-1$ chooses θ_{l-1}^1, $l+1$ chooses θ_{l+1}^1. In this case, θ_l^1 simply denotes the mentioned unique payoff type of l.

To sum up, the set of histories H of Γ consists of terminal histories $(\theta^{\mathrm{I}}, (z_i), (h^i), (k_i^0, k_i))$ and their initial subhistories, where we let the set of actions A be the union of all actions described in Stages I-IV. About every choice of every player in the second and third stage, as little as possible is revealed to the other players. E.g., while j may realize in Stage III that i chose *some* $z_i \neq r$ in Stage II (player j can infer this if j observes that some third-stage Substage $l \neq i$ is played), j will not learn before the end of the mechanism *which* $z_i \in Y_i^F(\theta_{-i}^{\mathrm{I}})$ agent i chose. As a second example, all Substage i choices of agent i in Stage III remain private until the end of the mechanism. The set of i's second-stage information sets is

$$\mathbb{H}_i^{\mathrm{II}} = \bigcup_{\theta' \in \Theta} \big\{\{(\theta', (z_j)_{j<i}) \in H\}\big\}$$

We refrain from formally defining the set $\mathbb{H}_i^{\mathrm{III},i}$ of i's third-stage Substage-i information sets, the set $\mathbb{H}_i^{\mathrm{III},-i}$ of i's other third-stage information sets and the set $\mathbb{H}_i^{\mathrm{IV}}$ of i's fourth-stage information sets.

To define the outcome function C of Γ, define $C^{\mathrm{II}}(\theta^{\mathrm{I}}, (z_i), (k_i^0))$ and $C^{\mathrm{III}}(\theta^{\mathrm{I}}, (h^i), (k_i))$ as in Figure 2, and then let

$$C(\theta^{\mathrm{I}}, (z_i), (h^i), (k_i^0, k_i)) = \frac{1}{2} C^{\mathrm{II}}(\theta^{\mathrm{I}}, (z_i), (k_i^0)) + \frac{1}{2} C^{\mathrm{III}}(\theta^{\mathrm{I}}, (h^i), (k_i)).^{13}$$

The set of actions A, the set of histories H, the set of i's information sets $\mathbb{H}_i = \{\mathcal{H}_i^{\mathrm{I}}\} \cup \mathbb{H}_i^{\mathrm{II}} \cup \mathbb{H}_i^{\mathrm{III},i} \cup \mathbb{H}_i^{\mathrm{III},-i} \cup \mathbb{H}_i^{\mathrm{IV}}$ for each $i \in \mathcal{I}$, the outcome function C, and the (obvious) player function P fully describe the mechanism Γ. In Γ every agent has only finitely many information sets and at most countably many actions available at each information set. Hence both A and S_i for all i are countable sets and Γ is a countable mechanism (as defined in Subsection 2.2). We now prove that Γ wr-implements f.

Step 1. If $(s_i, \theta_i) \in W_i^1$, $\theta^{\mathrm{I}} \in \Theta$ and $\mathcal{H}^{\mathrm{II}} = \{(\theta^{\mathrm{I}}, (z_j)_{j<i}) \in H\} \in \mathbb{H}_i(s_i)$, then $s_i(\mathcal{H}^{\mathrm{II}}) = r$.

Proof: If $\mathcal{H}^{\mathrm{II}} = \{(\theta^{\mathrm{I}}, (z_j)_{j<i}) \in H\}$ is a second-stage information set in $\mathbb{H}_i(s_i)$ then $s_i(\mathcal{H}_1^{\mathrm{I}}) = \theta_i^{\mathrm{I}}$ and i's available actions at $\mathcal{H}^{\mathrm{II}}$ are r and every $z_i \in Y_i^F(\theta_{-i}^{\mathrm{I}})$. Suppose that $s_i(\mathcal{H}^{\mathrm{II}}) = z_i \in Y_i^F(\theta_{-i}^{\mathrm{I}})$. Let μ_i be a CPS with respect to which s_i is sequentially rational for θ_i. Then for all $\mathcal{H} \in \mathbb{H}_i(s_i)$, s_i maximizes $U_i^{\mu_i}(\cdot, \theta_i, \mathcal{H})$. In particular, $U_i^{\mu_i}(s_i, \theta_i, \mathcal{H}^{\mathrm{II}}) \geq U_i^{\mu_i}(s_i', \theta_i, \mathcal{H}^{\mathrm{II}})$ for all $s_i' \in S_i(\mathcal{H}^{\mathrm{II}})$, that is, s_i is optimal in the case that $-i$ announces θ_{-i}^{I} in the first stage.

Recall that if $h = \big(\theta^{\mathrm{I}}, \big(z_i, (z_j)_{j \neq i}\big), (h^l)_{l \in \mathcal{I}}, (k_l^0, k_l)_{l \in \mathcal{I}}\big)$ ends up being the terminal history

[13] The definition of C implies that every agent i's fourth-stage choice k_i^0 is irrelevant (in that it does not influence Γ's outcome) if i chose $z_i = r$ in Stage II. Similarly, the choice of k_i^m, $m \geq 1$, is irrelevant if third-stage Substage i was not played, if $K_i^m = \{1\}$ or if j_i ended Substage i by playing $\neg c$ before i could choose K_i^m. For notational simplicity, we nonetheless let i choose the entire tuple (k_i^0, k_i) in these cases.

(where $(z_j)_{j\neq i}$, $(h^l)_{l\in\mathcal{I}}$ and $(k_l^0, k_l)_{l\in\mathcal{I}}$ are such that $h \in H$), then the outcome of Γ is

$$C(h) = \frac{1}{2I}\Big[\frac{1}{k_i^0}\bar{y}_{\theta^{\mathrm{I}}_{-i}} + \Big(1 - \frac{1}{k_i^0}\Big)z_i + \sum_{j\neq i} C_j^{\mathrm{II}}(\theta^{\mathrm{I}}, z_j, k_j^0)\Big] + \frac{1}{2}C^{\mathrm{III}}(\theta^{\mathrm{I}}, (h^l), (k_l)).$$

Every strategy s_i' that prescribes some element of $Y_i^F(\theta^{\mathrm{I}}_{-i})$ (potentially different from z_i) at $\mathcal{H}^{\mathrm{II}}$ and some actions in Stage IV (potentially different from the fourth-stage actions of s_i) but otherwise equals s_i is in $S_i(\mathcal{H}^{\mathrm{II}})$. Moreover, for every $j \neq i$, every such s_i' admits the same information sets $\mathcal{H} \in \mathbb{H}_j$ as s_i, as no $j \neq i$ can observe or deduce before the end of the mechanism which element i chooses from $Y_i^F(\theta^{\mathrm{I}}_{-i})$ in Stage II, and all agents' Stage IV choices are simultaneous. Thus, the third-stage subhistory expected by i at $\mathcal{H}^{\mathrm{II}}$ is the same whether she plays s_i or some such s_i'. In fact, the only effect that playing some such s_i' instead of s_i has on the outcome of Γ is on the term $\frac{1}{2I}\Big[\frac{1}{k_i^0}\bar{y}_{\theta^{\mathrm{I}}_{-i}} + \Big(1 - \frac{1}{k_i^0}\Big)z_i\Big]$.

Since by (6)

$$E_{marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H}^{\mathrm{II}})}u_i(y,\theta) > E_{marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H}^{\mathrm{II}})}u_i(\bar{y}_{\theta^{\mathrm{I}}_{-i}},\theta) \quad \text{for some } y \in Y_i^F(\theta^{\mathrm{I}}_{-i}),$$

the optimality of s_i at $\mathcal{H}^{\mathrm{II}}$ implies that $E_{marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H}^{\mathrm{II}})}u_i(z_i,\theta) > E_{marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H}^{\mathrm{II}})}u_i(\bar{y}_{\theta^{\mathrm{I}}_{-i}},\theta)$ and that $k_i^0 > 1$ for at least some fourth-stage information set that, given $\mu_i(\cdot|\mathcal{H}^{\mathrm{II}})$, i expects will be reached with strictly positive probability. Let $\mathcal{H}^{\mathrm{IV}} \in \mathbb{H}_i^{\mathrm{IV}}(s_i)$ be one such information set, and suppose $s_i(\mathcal{H}^{\mathrm{IV}}) = (k_i^0, k_i)$. Then the strategy s_i' that prescribes $(k_i^0 + 1, k_i)$ at $\mathcal{H}^{\mathrm{IV}}$ but otherwise equals s_i promises a higher expected utility to θ_i at $\mathcal{H}^{\mathrm{II}}$ than s_i. Contradiction.

Step 2. If $(s_i, \theta_i) \in W_i^1$ and $s_i(\mathcal{H}_i^{\mathrm{I}}) = \theta_i^{\mathrm{I}}$, then $\theta_i \notin N_i(\theta_i^{\mathrm{I}})$.

Proof: Suppose that $(s_i, \theta_i) \in W_i^1$ and $s_i(\mathcal{H}_i^{\mathrm{I}}) = \theta_i^{\mathrm{I}}$, but that there exists an m such that $\theta_i = \theta_i^m \in N_i(\theta_i^{\mathrm{I}}) = \{\theta_i^1, \ldots, \theta_i^{n_i}\}$. Fix an arbitrary $\theta^{\mathrm{I}}_{-i} \in \Theta_{-i}$. Since $i \in \mathcal{N}(\theta^{\mathrm{I}})$, there is a third-stage Substage i which is admitted by the first-stage announcements θ^{I} and by i's second-stage choice (which by Step 1 is) $s_i(\{(\theta^{\mathrm{I}}, (z_j)_{j<i}) \in H\}) = r$, and which is triggered if all $l \neq i$ fail to play r in the second stage. Let $\mathcal{H}^{\mathrm{III}} \in \mathbb{H}_i^{\mathrm{III},i}(s_i)$ be i's m-th information set of this Substage i. Since $(s_i, \theta_i^m) \in W_i^1$, there is a CPS $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ that rationalizes s_i for θ_i^m, that is, such that $(s_i, \theta_i^m) \in \rho_i(\mu_i)$. Define $\chi_i \in \Delta(supp(\psi_i'(\theta_i^{\mathrm{I}}, \theta_i^m)) \times \Theta_{-i})$ by

$$\chi_i(\theta'_{-i}, \theta_{-i}) = \sum_{\substack{s_{-i}\in S_{-i}\text{: } \forall j\neq i,\ s_j \text{ prescribes } \theta'_j \text{ at } j\text{'s first information set following } i\text{'s choice at } \mathcal{H}^{\mathrm{III}}}} \mu_i((s_{-i}, \theta_{-i})|\mathcal{H}^{\mathrm{III}})$$

for all $(\theta'_{-i}, \theta_{-i}) \in supp(\psi_i'(\theta_i^{\mathrm{I}}, \theta_i^m)) \times \Theta_{-i}$. Recall that i's choice at $\mathcal{H}^{\mathrm{III}}$ cannot be observed by any $j \neq i$ and thus cannot influence any $(j \neq i)$'s future actions. Since there is a robust $(\theta_i^{\mathrm{I}}, \theta_i^m)$-preference reversal, there exists a $q \in \{1, \ldots, p(\theta_i^{\mathrm{I}}, \theta_i^m)\}$ such that (9) (with $(\theta_i', \theta_i) = (\theta_i^{\mathrm{I}}, \theta_i^m)$) holds for χ_i. Since $(s_i, \theta_i^m) \in \rho_i(\mu_i)$, action $s_i(\mathcal{H}^{\mathrm{III}}) = (q_i^m, K_i^m) \in \{1, \ldots, p(\theta_i^{\mathrm{I}}, \theta_i^m)\} \times$

$\{\{1\},\{2,3,\ldots\}\}$ must thus satisfy $E_{\chi_i} u_i(x^{q_i^m}(\theta_i^{\mathrm{I}},\theta_i^m)(\theta'_{-i}),\theta) > E_{\chi_i} u_i(y(\theta_i^{\mathrm{I}},\theta_i^m)(\theta'_{-i}),\theta)$ and $K_i^m = \{2,3,\ldots\}$, or equivalently, $U_i^{\mu_i}(s_i,\theta_i,\mathcal{H}^{\mathrm{III}}) > U_i^{\mu_i}(s'_i,\theta_i,\mathcal{H}^{\mathrm{III}})$ for all $s'_i \in S_i(\mathcal{H}^{\mathrm{III}})$ such that s'_i equals s_i except that $s'_i(\mathcal{H}^{\mathrm{III}}) \in \{1,\ldots,p(\theta_i^{\mathrm{I}},\theta_i^m)\} \times \{\{1\}\}$. Let $\mathcal{H}^{\mathrm{IV}}$ be some fourth-stage information set that, given $\mu_i(\cdot|\mathcal{H}^{\mathrm{III}})$, i expects to reach with strictly positive probability and say that at this information set, s_i prescribes $s_i(\mathcal{H}^{\mathrm{IV}}) = (k_i^0,k_i^1,\ldots,k_i^{n_i})$. Then i prefers to play s'_i rather than s_i at $\mathcal{H}^{\mathrm{III}}$, where s'_i equals s_i except that

$$s'_i(\mathcal{H}^{\mathrm{IV}}) = (k_i^0,k_i^1,\ldots,k_i^{m-1},k_i^m+1,k_i^{m+1},\ldots,k_i^{n_i}).$$

Contradiction.

Step 3. For all $i \in \mathcal{I}$ and $\theta_i \in \Theta_i$, let

$$\begin{aligned} S_i(\theta_i) = \big\{ s_i \in S_i : {} & s_i(\mathcal{H}_i^{\mathrm{I}}) = \theta_i, \\ & s_i(\mathcal{H}^{\mathrm{II}}) = r \text{ for all } \mathcal{H}^{\mathrm{II}} \in \mathbb{H}_i^{\mathrm{II}}(s_i) \text{ and} \\ & s_i(\mathcal{H}^{\mathrm{III}}) \in \mathbb{N} \times \{\{1\}\} \text{ for all } \mathcal{H}^{\mathrm{III}} \in \mathbb{H}_i^{\mathrm{III},i}(s_i) \big\}. \end{aligned}$$

Then $s_i \in S_i(\theta_i)$ implies $(s_i,\theta_i) \in W_i^\infty$.

Proof: For all $\theta_{-i} \in \Theta_{-i}$, let $S_{-i}(\theta_{-i}) = \prod_{j\neq i} S_j(\theta_j)$. Let $\bar{\theta}_i \in \Theta_i$ and $\bar{s}_i \in S_i(\bar{\theta}_i)$ and let μ_i be a CPS such that

- for all $\mathcal{H} \in \bar{\mathbb{H}}_i(\{\bar{s}_i\} \times \bigcup_{\theta_{-i}\in\Theta_{-i}} S_{-i}(\theta_{-i}))$ (that is, for $\{\emptyset\}$, $\mathcal{H}_i^{\mathrm{I}}$, all $\mathcal{H} \in \mathbb{H}_i^{\mathrm{II}}(\bar{s}_i)$ and all $\mathcal{H} \in \mathbb{H}_i^{\mathrm{IV}}(\bar{s}_i)$ that immediately follow a Stage II information set because no Stage III substage was played), there are $\theta_{-i}^{\mathcal{H}} \in \Theta_{-i}$ and $s_{-i}^{\mathcal{H}} \in S_{-i}(\theta_{-i}^{\mathcal{H}})$ such that $\mu_i((s_{-i}^{\mathcal{H}},\theta_{-i}^{\mathcal{H}})|\mathcal{H}) = 1$, and
- for all $\mathcal{H}^{\mathrm{III}} \in \mathbb{H}_i^{\mathrm{III},i}(\bar{s}_i)$ and all $m \in \mathbb{N}$, if $\mathcal{H}^{\mathrm{III}}$ is i's m-th substage i information set that follows the first-stage announcements $\theta^{\mathrm{I}} \in \Theta$ such that $\theta_i^{\mathrm{I}} = \bar{\theta}_i$, and if θ_i^m is the m-th element of $N_i(\bar{\theta}_i)$, then $marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H}^{\mathrm{III}}) = \psi'_i(\bar{\theta}_i,\theta_i^m)$ and, for all $(s_{-i},\theta_{-i}) \in \Sigma_{-i}$, $\mu_i((s_{-i},\theta_{-i})|\mathcal{H}^{\mathrm{III}}) > 0$ implies that for each $j \neq i$, 1) s_j prescribes θ_j ("truth-telling") at j's information set that "immediately follows" $\mathcal{H}^{\mathrm{III}}$ and 2) if $m < n_i$ and $j = (i+1) \bmod I$ then s_j prescribes $\neg c$ the next time in Substage i that j decides between c and $\neg c$.

Note that we can construct such a CPS μ_i by starting from the "root" of the mechanism and working our way to the "leaves." We start by letting $\mu_i(\cdot|\{\emptyset\}) = \delta(s_{-i}^{\{\emptyset\}},\theta_{-i}^{\{\emptyset\}})$ for some $\theta_{-i}^{\{\emptyset\}}$ and some $s_{-i}^{\{\emptyset\}} \in S_{-i}(\theta_{-i}^{\{\emptyset\}})$. By Bayesian updating, this pins down $\mu_i(\cdot|\mathcal{H})$ as $\mu_i(\cdot|\mathcal{H}) = \delta(s_{-i}^{\{\emptyset\}},\theta_{-i}^{\{\emptyset\}})$, for all $\mathcal{H} \in \mathbb{H}_i(s_{-i}^{\{\emptyset\}})$. Next, we consider all information sets $\mathcal{H}'$ for which we have not defined $\mu_i(\cdot|\mathcal{H}')$ yet and that are an immediate successor to an information set $\mathcal{H}$ for which we already defined $\mu_i(\cdot|\mathcal{H})$. Every such $\mathcal{H}'$ is a surprise, and we can simply let $\mu_i(\cdot|\mathcal{H}')$ equal some probability measure that satisfies every of the two bullet points listed above (at most one of

which applies to $\mathcal{H}'$). Bayesian updating pins down $\mu_i(\cdot|\mathcal{H}'')$ at every $\mathcal{H}''$ that is reached with positive probability according to $\mu_i(\cdot|\mathcal{H}')$. Finally, we iterate the described procedure by next defining $\mu_i(\cdot|\mathcal{H}''')$ for every $\mathcal{H}'''$ for which $\mu_i(\cdot|\mathcal{H}''')$ is yet undefined and such that $\mathcal{H}'''$ is an immediate successor to an information set $\mathcal{H}$ for which we already defined $\mu_i(\cdot|\mathcal{H})$, and so on.

We claim that $\bar{s}_i \in r_i(\bar{\theta}_i, \mu_i)$.

To see this, consider $\mathcal{H} \in \mathbb{H}_i(\{\bar{s}_i\} \times \bigcup_{\theta_{-i} \in \Theta_{-i}} S_{-i}(\theta_{-i}))$. We have

$$U_i^{\mu_i}(\bar{s}_i, \bar{\theta}_i, \mathcal{H}) = u_i(f(\bar{\theta}_i, \theta_{-i}^{\mathcal{H}}), \bar{\theta}_i, \theta_{-i}^{\mathcal{H}}) \geq U_i^{\mu_i}(s_i, \bar{\theta}_i, \mathcal{H})$$

for every $s_i \in S_i(\mathcal{H})$, as

$$\begin{aligned}\zeta(s_i, s_{-i}^{\mathcal{H}}) \in \Bigg\{ & (1-\alpha-\beta) f(\theta_i^{\mathrm{I}}, \theta_{-i}^{\mathcal{H}}) + \alpha \left(\frac{1}{k_i^0} \bar{y}_{\theta_{-i}^{\mathcal{H}}} + \left(1 - \frac{1}{k_i^0}\right) z_i \right) \\ & \qquad + \beta \left((1-\varepsilon) \bar{y}_{\theta_{-i}^{\mathcal{H}}} + \varepsilon \sum_{l=1}^{n_j} q(l) y(\theta_j^{\mathcal{H}}, \theta_j^l)(\theta_i^l) \right) : \\ & I = 2 \Rightarrow (\alpha, \beta) \in \left\{(0,0), \left(\tfrac{1}{4}, 0\right), \left(\tfrac{1}{4}, \tfrac{1}{4}\right)\right\}, I > 2 \Rightarrow (\alpha, \beta) \in \left\{(0,0), \left(\tfrac{1}{2I}, 0\right)\right\}, \\ & \theta_i^{\mathrm{I}} \in \Theta_i, z_i \in Y_i^F(\theta_{-i}^{\mathcal{H}}), k_i^0 \geq 1, i = (j+1) \mod I, \forall l : \theta_i^l \in supp(\psi_j'(\theta_j^{\mathcal{H}}, \theta_j^l)), \\ & n_j = \#N_j(\theta_j^{\mathcal{H}}) = \#\{\theta_j^1, \ldots, \theta_j^{n_j}\}, \sum_{l=1}^{n_j} q(l) = 1, \forall l : q(l) \geq 0 \Bigg\}\end{aligned}$$

and $\bar{\theta}_i$ prefers $f(\bar{\theta}_i, \theta_{-i}^{\mathcal{H}})$ over $f(\theta_i^{\mathrm{I}}, \theta_{-i}^{\mathcal{H}})$ (the social choice function f is epIC) and over every element of $Y_i(\theta_{-i}^{\mathcal{H}})$ (by construction of the reward set), and because ε satisfies (7).

In order to establish that $\bar{s}_i$ is sequentially rational for $\bar{\theta}_i$ given μ_i, we need to verify that $\bar{s}_i$ is optimal at every information set $\mathcal{H}$ admitted by $\bar{s}_i$, even if $\mathcal{H}$ is not admitted by sequentially rational play of $-i$. Therefore, also consider the third-stage Substage-i information sets that i may encounter if she follows $\bar{s}_i$. Let $\theta \in \Theta$ denote some first-stage announcements such that $\theta_i = \bar{\theta}_i$, and say that $\mathcal{H}^{\mathrm{III}} \in \mathbb{H}_i^{\mathrm{III},i}(\bar{s}_i)$ is the m-th information set that i encounters in the Substage i that follows θ. Let θ_i^m denote the m-th element of $N_i(\bar{\theta}_i)$. By definition of μ_i, 1) $marg_{\Theta_{-i}} \mu_i(\cdot|\mathcal{H}^{\mathrm{III}}) = \psi_i'(\bar{\theta}_i, \theta_i^m)$ and i believes 2) that $-i$ will truthfully announce their payoff types immediately following $\mathcal{H}^{\mathrm{III}}$ (independently of i's choice at $\mathcal{H}^{\mathrm{III}}$) and 3) that j_i will end Substage i at the next opportunity, unless $m = n_i$ and Substage i ends anyway after $-i$'s payoff type announcements. Moreover, 4) i's action $s_i(\mathcal{H}^{\mathrm{III}})$ remains private until the end of the mechanism, and hence i cannot believe that her choice of $s_i(\mathcal{H}^{\mathrm{III}})$ influences $-i$'s remaining choices (this matters because i knows that Substage i is played and thus that for all $l \neq i$, agent l's choice of k_l^0 in Stage IV will affect the mechanism's outcome). Thus by (9), $s_i(\mathcal{H}^{\mathrm{III}}) \in \{1, \ldots, p(\bar{\theta}_i, \theta_i^m)\} \times \{\{1\}\}$ is optimal at $\mathcal{H}^{\mathrm{III}}$ by construction.

Every $\mathcal{H}^{\mathrm{III}} \in \mathbb{H}_i^{\mathrm{III},-i}$, that is, every third-stage information set from one of the Substages

$j \neq i$, is not admitted by $\bar{s}_i$ because if i plays r in Stage II then all Substages $j \neq i$ will be skipped. Hence we do not need to check optimality at any $\mathcal{H}^{\text{III}} \in \mathbb{H}_i^{\text{III},-i}$. At every fourth-stage information set $\mathcal{H}^{\text{IV}}$ admitted by $\bar{s}_i$, all available actions lead to the same outcome because $\bar{s}_i$ prescribed r in Stage II (which makes the choice of k_i^0 irrelevant) and $K_i^m = \{1\}$ at every Substage i information set of i (which forces the choice $k_i^m = 1$ in all cases in which k_i^m enters the outcome function). Hence the optimality of $\bar{s}_i$ at $\mathcal{H}^{\text{IV}}$ is trivial.

Because $\bar{s}_i \in r_i(\bar{\theta}_i, \mu_i)$ we have $\bar{s}_i \in Q_i^1(\bar{\theta}_i)$. By a symmetric argument, $\bigcup_{\theta_{-i} \in \Theta_{-i}} (S_{-i}(\theta_{-i}) \times \{\theta_{-i}\}) \subseteq W_{-i}^1$. Hence $\mu_i \in \Pi_i^1$ and $\bar{s}_i \in Q_i^2(\bar{\theta}_i)$. And so on. By transfinite induction, $\bar{s}_i \in Q_i^\infty(\bar{\theta}_i)$.

Step 4. If $s_i \in Q_i^\infty(\theta_i)$ and $s_i(\mathcal{H}_i^1) = \theta_i^{\text{I}}$, then $f(\theta_i, \theta_{-i}) = f(\theta_i^{\text{I}}, \theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$.

Proof: To see this, consider the deception β such that for all $i \in \mathcal{I}$ and all $\theta_i \in \Theta_i$,

$$\beta_i(\theta_i) = \{\theta_i^{\text{I}} \in \Theta_i : \exists s_i \in Q_i^\infty(\theta_i), s_i(\mathcal{H}_i^{\text{I}}) = \theta_i^{\text{I}}\}$$

(Step 3 ensures that $\theta_i \in \beta_i(\theta_i)$ for all i and θ_i, and thus that β is indeed a deception.) If β is acceptable then Step 4's claim is true. Suppose by contradiction that β is unacceptable. Then by the hypothesis that f is dr-monotone, β is d-refutable. Let $i = i(\beta)$, $\theta_i = \theta_i(\beta)$ and $\theta_i^{\text{I}} = \theta_i'(\beta) \in \beta_i(\theta_i)$. Since by the definition of β, there exists a $s_i \in Q_i^\infty(\theta_i) \subseteq Q_i^1(\theta_i)$ such that $s_i(\mathcal{H}_i^1) = \theta_i^{\text{I}}$, by Step 2, there is no robust $(\theta_i^{\text{I}}, \theta_i)$-preference reversal. Therefore, the d-refutability of β implies that for each $\theta'_{-i} \in \Theta_{-i}$ there is a finite subset $Y_i^F(\theta'_{-i})$ of $Y_i(\theta'_{-i})$ such that for all $\psi_i \in \Delta(\Theta_{-i})$ with $\psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1$, (8) holds.

Let $s_i \in Q_i^\infty(\theta_i)$ be such that $s_i(\mathcal{H}_i^{\text{I}}) = \theta_i^{\text{I}}$ and $\mu_i \in \Pi_i^\infty$ be such that s_i is a sequential best response for θ_i against μ_i.[14] Choose a $\theta_{-i}^{\text{I}} \in \Theta_{-i}$ such that $\mu_i(\bar{S}_{-i}(\theta_{-i}^{\text{I}}) \times \Theta_{-i} | \mathcal{H}_i^{\text{I}}) > 0$, where

$$\bar{S}_{-i}(\theta_{-i}^{\text{I}}) = \left\{s_{-i} \in S_{-i} : s_j(\mathcal{H}_j^{\text{I}}) = \theta_j^{\text{I}} \text{ for all } j \neq i\right\}.$$

[14] Such a μ_i exists because our definition by transfinite recursion of W^∞ implies that $W_i^\infty = \rho_i(\Pi_i^\infty)$ for all $i \in \mathcal{I}$ (see proof of Lemma 3 in Appendix B). If we had defined weak rationalizability as W^{ω_0}, the existence of such a μ_i would not be immediate (because in some mechanisms $W_i^{\omega_0} \neq \rho_i(\bigcap_{n \in \mathbb{N}} \Pi_i^n)$ for some $i \in \mathcal{I}$; e.g., in the mechanism of Example A.1 in Appendix A,

$$\{-1, 0\} \times \Theta_i = W_i^{\omega_0} \neq \rho_i(\bigcap_{n \in \mathbb{N}} \Pi_i^n) = \rho_i(\Delta(\{-1, 0\})) = \{-1\} \times \Theta_i$$

for all $i \in \mathcal{I}$), but would still follow because in the mechanism Γ of the proof of Proposition 3, the iterated elimination of never-best sequential best responses converges in finitely many rounds. To see that there exists a $k \in \mathbb{N}$ such that $W^{k'} = W^{\omega_0} = W^\infty$ for all $k' \geq k$, recall that the only information sets with more than finitely many actions are the Stage IV information sets. By Step 1, if all players are sequentially rational, then all players choose r in Stage II (implying e.g. that all Substages i will be skipped). At every information set admitted by W^1 and with respect to every $\mu_i \in \Pi_i^1$, this renders i indifferent between all Stage IV actions. Therefore, after one round of elimination, only finitely many relevant "equivalence classes" of strategies remain.

For every $\theta_{-i} \in \Theta_{-i}$, let

$$\psi_i(\theta_{-i}) = \frac{\mu_i(\bar{S}_{-i}(\theta^{\mathrm{I}}_{-i}) \times \{\theta_{-i}\}|\mathcal{H}^{\mathrm{I}}_i)}{\mu_i(\bar{S}_{-i}(\theta^{\mathrm{I}}_{-i}) \times \Theta_{-i}|\mathcal{H}^{\mathrm{I}}_i)}.$$

Then $\psi_i(\beta^{-1}_{-i}(\theta^{\mathrm{I}}_{-i})) = 1$. Consider i's information set $\mathcal{H}^{\mathrm{II}} = \{(\theta^{\mathrm{I}}, (z_j)_{j<i}) \in H\}$. Note that $\psi_i = marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H}^{\mathrm{II}})$. By (8) for θ^{I}_{-i} and ψ_i

$$E_{\psi_i}u_i(x^{(\beta,\theta^{\mathrm{I}}_{-i},\psi_i)}, \theta_i, \theta_{-i}) > E_{\psi_i}u_i(f(\theta^{\mathrm{I}}), \theta_i, \theta_{-i}),$$

and there is a strategy $s'_i \in S_i(\mathcal{H}^{\mathrm{II}})$ which provides θ_i with a strictly higher expected utility at $\mathcal{H}^{\mathrm{II}}$ than s_i (which by Step 1 prescribes r at $\mathcal{H}^{\mathrm{II}}$): the choice of θ^{I}_{-i} guarantees that $\mu_i(\cdot|\mathcal{H}^{\mathrm{II}})$ is a Bayesian update of $\mu_i(\cdot|\mathcal{H}^{\mathrm{I}}) = \mu_i(\cdot|\{\emptyset\})$. Hence if $I \geq 3$, at $\mathcal{H}^{\mathrm{II}}$, agent i expects all $j \neq i$ to play r in Stage II and thus that Stage III will be skipped no matter which action she chooses at $\mathcal{H}^{\mathrm{II}}$. Hence if $I \geq 3$, it suffices to let s'_i prescribe $x^{(\beta,\theta^{\mathrm{I}}_{-i},\psi_i)}$ at $\mathcal{H}^{\mathrm{II}}$ and k^0_i in Stage IV for k^0_i large enough. If $I = 2$, agent i expects that playing $x^{(\beta,\theta^{\mathrm{I}}_{-i},\psi_i)}$ triggers Stage III Substage $j \neq i$, and we additionally require that s'_i specifies to end Substage $j \neq i$ immediately by choosing $\neg c$ at the first opportunity. Contradiction to s_i being sequentially rational for θ_i with respect to μ_i.

Step 5. By Steps 1 and 4, for every $(s, \theta) \in \Sigma$, if s is weakly rationalizable for θ then $C(\zeta(s)) = f(\theta)$. The proof of Step 3 implies that for all $i \in \mathcal{I}$ and $\theta_i \in \Theta_i$, $\emptyset \neq S_i(\theta_i) \subseteq Q^{\infty}_i(\theta_i)$, and that for all $i \in \mathcal{I}$, $\theta \in \Theta$ and $s_{-i} \in S_{-i}(\theta_{-i})$, there exist $s_i \in S_i(\theta_i)$ and $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ such that $\mu_i((s_{-i}, \theta_{-i})|\{\emptyset\}) = 1$ and $s_i \in r_i(\theta_i, \mu_i)$. Hence Γ wr-implements f. □

This completes the proof of Proposition 3. Note that, as mentioned in Footnote 14, the iterated elimination of never-best sequential best responses converges in finitely many rounds.

4 WR-Implementation and Robust wPBE-Implementation

In this section we provide a foundation for wr-implementation by proving its equivalence to robust wPBE-implementation. In broad terms, we thus demonstrate that a well-known motivation from the static case extends to dynamic mechanisms: BM show that rationalizable implementation by static mechanisms (as defined by BM) is *almost* equivalent to and thus motivated by robust wPBE-implementation by static mechanisms. Our equivalence analogously motivates our interest in wr-implementation. A more detailed look reveals that our result does strictly speaking not generalize BM's. This is because the notion of wr-implementation by static mechanisms slightly differs from rationalizable implementation and is *exactly* equivalent to robust wPBE-implementation by static mechanisms (see Subsection 4.4). Our result thus improves upon BM's while extending it to dynamic mechanisms.

Viewed differently, this section provides an implementation theory counterpart to the game theoretic result that for every payoff type, the union of weakly perfect Bayesian equilibrium strategies across all type spaces corresponds to the set of weakly rationalizable strategies (Battigalli, 1999; see also Battigalli and Siniscalchi, 2003).[15] Deriving our result requires additional work essentially because A) to achieve implementation, we need to guarantee that the set of wPBE is non-empty for all type spaces, an aspect that the game theoretic equivalence is silent about, and B) while the game-theoretic result has been established for "simple" dynamic games, we state our result for a general class of mechanisms that includes mechanisms that are not "simple."

4.1 An Equivalence Result for Countable Mechanisms

We define robustness in the spirit of the belief-free implementation literature and require implementation across all type spaces. A type space consists of, for each agent i, a set T_i of (epistemic) types, a function $\hat{\theta}_i$ that associates with each type t_i a payoff type θ_i and a function τ_i that associates with each type t_i a belief on T_{-i} (and thus implicitly, also on Θ_{-i}). Each type space represents a "situation" in which the mechanism may be played, describing the beliefs that every payoff type of every agent may hold at the beginning of the mechanism.

Definition 9 *A type space is a tuple $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i\in\mathcal{I}}$ such that for every $i \in \mathcal{I}$, T_i is a non-empty topological space of player i's types, $\hat{\theta}_i : T_i \to \Theta_i$ is measurable and $\tau_i : T_i \to \Delta(T_{-i})$.*[16,17]

We follow Battigalli (1999) in defining weak perfect Bayesian equilibrium (save for the measurability assumption discussed in Footnote 17).

[15] Similar game theoretic results have been put forward for various solution concepts. E.g., Brandenburger and Dekel (1987) relate a posteriori equilibrium and rationalizability in finite games of complete information; Battigalli and Siniscalchi (2003) relate Bayesian equilibrium consistent with a set of first-order beliefs Δ and Δ-rationalizability in finite games of incomplete information; BM relate interim equilibrium and belief-free rationalizability in countable games of incomplete information; Penta (2015) relates interim perfect equilibrium to (belief-free) backwards rationalizability in compact multi-stage games of incomplete information.

[16] With minor modifications to the proofs, all results of Section 4 are also true for the following alternative definition. A type space is a tuple $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i\in\mathcal{I}}$ such that for every $i \in \mathcal{I}$, T_i is a non-empty measurable space of player i's types, $\hat{\theta}_i : T_i \to \Theta_i$ is measurable and $\tau_i : T_i \to \Delta(T_{-i})$, where we endow T_{-i} with the product σ-algebra. In this case we endow Σ_{-i} with the product σ-algebra $\mathcal{B}'_{-i}$ of the Borel σ-algebras on S_j, $j \neq i$, and Θ_j, $j \neq i$, instead of the Borel σ-algebra $\mathcal{B}_{-i}$ of the product topology on Σ_{-i} as we do outside of this footnote. Since $\mathcal{B}'_{-i} = \mathcal{B}_{-i}$ (in fact, $\mathcal{B}'_{-i} = \mathcal{B}_{-i} = 2^{\Sigma_{-i}}$) for every countable mechanism this change of the σ-algebra on Σ_{-i} is meaningless as long as we focus on countable mechanisms. However, this change becomes relevant in Subsection 4.2 as $\mathcal{B}'_{-i} \neq \mathcal{B}_{-i}$ for some general mechanisms. (In general, $\mathcal{B}'_{-i} \subseteq \mathcal{B}_{-i}$. But $\mathcal{B}'_{-i} = \mathcal{B}_{-i}$ if $I = 2$ or if S_j is a second-countable topological space for all $j \neq i$. If $I = 2$, this follows from the proof of Bogachev (2007, Lemma 6.4.2.(i)), who demonstrates that $\mathcal{B}_{-i}$ equals the product σ-algebra $\mathcal{B}''_{-i}$ of the Borel σ-algebra on S_{-i} with respect to the product topology on S_{-i} and the Borel σ-algebra $2^{\Theta_{-i}}$ on Θ_{-i}, and if S_j is second-countable for all $j \neq i$, by a related, standard argument.)

[17] As we will not explicitly construct belief hierarchies, we will not require measurability of τ_i or even introduce a σ-algebra on $\Delta(T_{-i})$ in our definition of a type space. Similarly, we will not require the belief maps g_i to be measurable in the upcoming definition of wPBE. (These omissions are irrelevant in countable type spaces, even if one is interested in explicitly constructing belief hierarchies.)

Definition 10 *Let Γ be a countable mechanism and $\mathcal{T}$ be a type space. An array $(b_i, g_i)_{i\in\mathcal{I}}$ of measurable functions $b_i : T_i \to S_i$ and functions $g_i : T_i \to \Delta^{\mathbb{H}_i}(\Sigma_{-i})$ is a weak perfect Bayesian equilibrium (wPBE) of Γ for $\mathcal{T}$ if for all $i \in \mathcal{I}$ and $t_i \in T_i$,*

- *(sequential rationality) $b_i(t_i) \in r_i(\hat{\theta}_i(t_i), g_i(t_i))$ and*
- *(consistency) for all $B_{-i} \subseteq \Sigma_{-i}$,*

$$g_i(t_i)(B_{-i}|\{\emptyset\}) = \tau_i(t_i)\{t_{-i} \in T_{-i} : (b_{-i}(t_{-i}), \hat{\theta}_{-i}(t_{-i})) \in B_{-i}\}.$$

While the definition of weak rationalizability (Definition 3) did not involve a type space, we defined wPBE with respect to a particular type space. That is, weak rationalizability is "explicitly belief-free," while wPBE depends on a description of the agents' belief hierarchies possible in the environment, as captured by the type space. A class of type spaces often used in applied models consists of "payoff type spaces" in which $T_i = \Theta_i$, $\hat{\theta}_i$ is the identity function and $\tau_i(\theta_i) = p(\cdot|\theta_i)$ for some common prior $p \in \Delta(\Theta)$. Other type spaces include belief hierarchies that are not necessarily derived from a a common prior. As common in the robust implementation literature, we reconcile the approach of describing beliefs via a type space with the explicitly belief-free approach by rendering the former approach implicitly belief-free by requiring implementation for all type spaces.

Definition 11 *Mechanism Γ robustly wPBE-implements social choice function f if for every type space $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i\in\mathcal{I}}$, 1) there exists a wPBE and 2) for every wPBE $(b_i, g_i)_{i\in\mathcal{I}}$ and every $t \in T$, $C(\zeta(b(t))) = f(\hat{\theta}(t))$.*

The following Corollary 1 states the characterization result of this section for countable mechanisms (such as the mechanism employed in our sufficiency result, Proposition 3). Corollary 1 will immediately follow from the more general Theorem 2 of Subsection 4.3.

Corollary 1 *Let Γ be a countable mechanism and f be a social choice function. Then Γ wr-implements f if and only if Γ robustly wPBE-implements f.*

Corollary 1 says that if a designer finds a countable mechanism Γ that wr-implements a desired social choice function, then the *same* mechanism also robustly wPBE-implements the social choice function, and vice versa. Obtaining such a mechanism by mechanism equivalence is stronger than obtaining the equality of the corresponding sets of implementable social choice functions. It is of potential importance for example if Γ satisfies additional desiderata beyond robustness, as it excludes the possibility that Γ wr-implements f but only a different mechanism Γ' (which may violate the additional desiderata) robustly wPBE-implements f.

4.2 General Definitions of Mechanisms and Related Concepts

Let us now formally define the type of mechanisms to which we will generalize Corollary 1 to in Subsection 4.3. With respect to wr-implementation, note that the proofs collected in Appendix B imply that permitting these type of mechanisms (instead of only focusing on countable mechanisms) does not weaken the necessary conditions.

Definition 12 *A mechanism is a tuple* $\Gamma = \langle A, H, (\mathbb{H}_i)_{i\in\mathcal{I}}, P, C\rangle$ *such that*

- A *is a* T_1 *topological space. We call elements of* A *actions.*[18]
- H *is a nonempty set of sequences*[19] *with codomain* A *such that*
  - *with* h *every initial subsequence of* h *is in* H *and*
  - *if every initial subsequence of a countably infinite sequence* h *is in* H*, then* $h \in H$*.*

  We call H *the set of histories. For every finite* $h \in H$*, we let* $A(h) = \{a \in A : (h, a) \in H\}$ *denote the set of actions available at* h *and endow it with the relative topology. We let* $T = \{h \in H : h$ *is countably infinite or* (h *is finite and* $A(h) = \emptyset$)$\}$ *be the set of terminal histories and call* $\emptyset \in H$ *the initial history. We write* $h' \preceq h$ *if* $h' \in H$ *is an initial subsequence of* $h \in H$*.*
- $P : H\backslash T \to \mathcal{I}$ *is a surjection. We call* P *the player function.*
- *for each* $i \in \mathcal{I}$*,* $\mathbb{H}_i$ *is a partition of* $\{h \in H\backslash T : P(h) = i\}$ *such that*
  - *for all* $\mathcal{H} \in \mathbb{H}_i$ *and all* $h, h' \in \mathcal{H}$*,* $A(h) = A(h')$*. Because of this property, we can write* $A(\mathcal{H})$ *for* $A(h)$ *for every* $h \in \mathcal{H}$*.*
  - *for all* $\mathcal{H} \in \mathbb{H}_i$ *and all* $h, h' \in H$*, if* $h \in \mathcal{H}$*,* $h' \preceq h$ *and* $h' \neq h$ *then* $h' \notin \mathcal{H}$*.*

  We call $S_i = \{s_i \in A^{\mathbb{H}_i} : \forall \mathcal{H} \in \mathbb{H}_i\,(s_i(\mathcal{H}) \in A(\mathcal{H}))\}$ *the set of strategies of* i*. The sets* $S_i(\mathcal{H})$ *for* $i \in \mathcal{I}$ *and* $\mathcal{H} \in \bigcup_{j\in\mathcal{I}} \mathbb{H}_j$ *and* $H((s_j)_{j\in\mathcal{J}})$ *for* $\mathcal{J} \subseteq \mathcal{I}$ *and* $(s_j)_{j\in\mathcal{J}} \in \prod_{j\in\mathcal{J}} S_j$ *etc. are defined as in Subsection 2.2. For every strategy profile* $s \in S$*, we let* $\zeta(s)$ *denote the terminal history induced by* s*.*
  We define a binary relation $\preceq$ *on* $\mathbb{H}_i$ *by writing* $\mathcal{H}' \preceq \mathcal{H}$ *if there are* $h' \in \mathcal{H}'$ *and* $h \in \mathcal{H}$

[18] Even if A is countable, the assumption of T_1 suffices for all of our results. That is, even though at the beginning of Section 2, for simplicity of exposition, we endowed every countable set with its discrete topology, all of our results continue to hold if we endow countable action sets A with T_1 topologies that are not discrete.

[19] A sequence is finite or countably infinite. A finite sequence h of length $n \in \mathbb{N}$ with codomain A is a function $h : \{1, \ldots, n\} \to A$. A countably infinite sequence h with codomain A is a function $h : \{1, 2, \ldots\} \to A$; its length is ∞. A finite sequence $g : \{1, \ldots, k\} \to A$ is an initial subsequence of a sequence h if the length of h is at least k and $g(l) = h(l)$ for all $l \in \{1, \ldots, k\}$. Note that $\emptyset$ (the unique finite sequence mapping $\{1, \ldots, 0\} = \emptyset$ to A) is an initial subsequence of every sequence with codomain A. For $h : \{1, \ldots, n\} \to A$ and $a \in A$, (h, a) denotes the finite sequence that maps $\{1, \ldots, n+1\}$ to A, has h as an initial subsequence and maps $n+1$ to a.

such that $h' \preceq h$*, and extend this relation to* $\bar{\mathbb{H}}_i = \mathbb{H}_i \cup \{\{\emptyset\}\}$ *(if necessary) by letting* $\{\emptyset\} \preceq \mathcal{H}$ *for all* $\mathcal{H} \in \bar{\mathbb{H}}_i$*.*

- $C : T \to Y$*. We call* C *the outcome function.*
- (perfect recall) *for all* $i \in \mathcal{I}$*,* $s_i \in S_i$ *and* $\mathcal{H} \in \mathbb{H}_i$*, if* $\mathcal{H} \cap H(s_i) \neq \emptyset$ *then* $\mathcal{H} \subseteq H(s_i)$*.*
- (no trivial decisions) *for all* $(h, a) \in H$ *with* h *finite there exists an action* $a' \neq a$ *such that* $(h, a') \in H$*.*
- (measurability) *for all* $i \in \mathcal{I}$ *and* $s_i \in S_i$*, the function* $C(\zeta(s_i, \cdot)) : S_{-i} \to Y, s_{-i} \mapsto C(\zeta(s))$ *is measurable.*[20]

The first five bullet points of Definition 12 imply that a mechanism is an extensive game form (see e.g. Osborne and Rubinstein (1994) for a definition of extensive game forms using the "history notation" we adapted here). Not every extensive game form is a mechanism, though. Definition 12 stipulates some mild extra conditions for an extensive game form to be called a mechanism. These have mostly technical motivations: We require the set A of actions to be a T_1 topological space to ensure that singletons $\{(s_{-i}, \theta_{-i})\}$ comprised of one strategy-payoff type profile are measurable subsets of Σ_{-i}. The requirement that the player function P is surjective simplifies notation.[21] The well-known conditions of perfect recall and no trivial decisions ensure that our description of players as Bayesian agents (made in Subsection 2.3) is sensible. And the measurability condition (which we use in the proof of Proposition 1) ensures that the agents' expectations are well-behaved in infinite mechanisms.

The class of (dynamic) mechanisms contains and is strictly larger than the class of static mechanisms. A mechanism is *static* if we can identify T with S and if each agent has exactly one information set. If $\mathcal{H}_i$ denotes i's single information set in a static mechanism Γ, then we can identify S_i with $A(\mathcal{H}_i)$, T with $\prod_i A_i(\mathcal{H}_i)$, and for each $s \in S$, $\zeta(s)$ with s. Assuming that $A = \bigcup_{i \in \mathcal{I}} A(\mathcal{H}_i)$ and that the topology on A is obvious, a static mechanism Γ is entirely specified by $(S_i)_{i \in \mathcal{I}}$ and C.

A mechanism is *countable* if A is countable and endowed with the discrete topology and S_i is countable for all $i \in \mathcal{I}$. Note that due to the no trivial decisions condition, the countability of S_i for all i implies that all histories of a countable mechanism are finite. That is, all countable mechanisms are of finite length.

We adopt the following, usual notational convention.

[20] Recall that as the product of the sets $A(\mathcal{H})$ for $\mathcal{H} \in \mathbb{H}_j$, S_j is endowed with the product topology. The set S_{-i}, in turn, is endowed with the product of the topologies on the sets S_j, $j \neq i$, and the corresponding Borel σ-algebra.

[21] P being surjective rules out the case that there is an "inactive" player $i \notin P(H \backslash T)$. This is not a substantive restriction as Definition 12 permits that i is "de facto inactive" in that none of her actions influence the outcome of the mechanism. Ruling out inactive players simplifies notation as it guarantees that every i has CPSs on $(\Sigma_{-i}, \bar{\mathbb{H}}_i)$ rather than $((\prod_{j \in \mathcal{J}, j \neq i} S_j \times \Theta_j) \times (\prod_{j \notin \mathcal{J}, j \neq i} \Theta_j), \bar{\mathbb{H}}_i)$ where $\mathcal{J}$ is the set of active players etc.

- **Convention (C).** If $(\Omega, \mathscr{F}, P)$ is a probability space and $F \subseteq \Omega$ is not measurable (i.e. $F \notin \mathscr{F}$), then every proposition containing the expression $P(F)$ is false.

In order for them to also apply to all uncountable mechanisms, we now generalize the definitions of Subsections 2.3, 2.4 and 4.1 as follows:

- **Subsection 2.3, including Definition 1 (CPS).** In the countable mechanism case, $\mathcal{B}_{-i}$ denoted the discrete σ-algebra $2^{\Sigma_{-i}}$. Now, we read $\mathcal{B}_{-i}$ to denote the Borel σ-algebra on Σ_{-i}.

- **Definition 2 (Sequential Rationality).** We say that both sides of (1) "make sense" if $u_i(C(\zeta(s_i, \cdot)), \theta_i, \cdot) : \Sigma_{-i} \to \mathbb{R}$ and $u_i(C(\zeta(s_i', \cdot)), \theta_i, \cdot) : \Sigma_{-i} \to \mathbb{R}$ are measurable with respect to the Borel σ-algebra $\mathcal{B}_{-i}$ completed with respect to $\mu_i(\cdot|\mathcal{H})$. Given a general mechanism, call a strategy $s_i \in S_i$ sequentially rational for payoff type $\theta_i \in \Theta_i$ of player i with respect to the beliefs $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ if for all $\mathcal{H} \in \mathbb{H}_i(s_i)$ and all $s_i' \in S_i(\mathcal{H})$, (1) holds and both sides of (1) "make sense."[22]

  In this paper we tacitly use the fact that if $\mu_i(\cdot|\mathcal{H})$, $\mathcal{H} \in \bar{\mathbb{H}}_i$, assigns all probability mass to finitely many mass points then $U_i^{\mu_i}(s_i, \theta_i, \mathcal{H})$ "makes sense" for all $s_i \in S_i$ and $\theta_i \in \Theta_i$.

- **Definition 3 (Weak Rationalizability).** Definition 3 applies verbatim to general mechanisms if we take into account the notational Convention (C) (which implies that $\mu_i(W_{-i}^{\alpha}|\{\emptyset\}) = 1$ is not satisfied if W_{-i}^{α} is not measurable).

- **Definition 10 (wPBE).** To generalize Definition 10, replace "countable mechanism" with "mechanism." Additionally, require the consistency condition only for all measurable (instead of for all) $B_{-i} \subseteq \Sigma_{-i}$. Finally, note that the notational Convention (C) implies that the consistency condition fails if for some measurable set $B_{-i} \subseteq \Sigma_{-i}$, the set $\{t_{-i} \in T_{-i} : (b_{-i}(t_{-i}), \hat{\theta}_{-i}(t_{-i})) \in B_{-i}\}$ is not measurable.

Definitions 4 (of wr-implementation), 5 (of d-refutability), 6 (of dr-monotonicity), 7 (of epIC), 8 (of the conditional NTI), 9 (of a type space) and 11 (of robust wPBE-implementation) continue to apply verbatim.

4.3 An Equivalence Result for General Mechanisms

After introducing the following technical assumption, we can generalize Corollary 1 to the general mechanisms from Definition 12.

$$\forall i \in \mathcal{I}, \alpha \in \text{Ord} : W_{-i}^{\alpha} \text{ is measurable} \tag{M}$$

[22] Since the Lebesgue integral is formally well-defined even for non-measurable functions and X and Θ are finite, both sides of (1) are well-defined and finite even for general mechanisms. We nonetheless add to Definition 2 the requirement that both sides of (1) "make sense" because for some non-measurable functions they might not have their usual interpretation.

Condition (M) guarantees that every agent i can assign a probability to the set W_{-i}^{α} of her opponents' weakly α-rationalizable strategy-payoff type profiles, for every ordinal number α.

Theorem 2 *Let Γ be a mechanism and f be a social choice function.*

(a) Γ wr-implements f if and only if Γ robustly wPBE-implements f and (M).

(b) Suppose the conditional NTI property is satisfied. Then f is wr-implementable if and only if f is robustly wPBE-implementable.

Theorem 2(b) asserts that in the class of social choice functions that satisfy the conditional NTI, the subsets of wr- and robustly wPBE-implementable social choice functions coincide. Theorem 2(a) strengthens this result to a mechanism by mechanism equivalence. This latter equivalence assumes a mechanism that satisfies (M), but applies even if the conditional NTI is violated.

We prove Theorem 2 in Appendix B by combining known techniques and new insights. To prove the "if" direction of Part (a), first, we show that W^{∞} corresponds to a wPBE of a particular type space. In establishing this auxiliary result for general mechanisms, our proof does not rely on various assumptions of related results in the literature (see the discussion preceding Lemma 3 in Appendix B for details) but instead adapts some arguments from Echenique (2005). Second, we prove a novel lemma that shows that robust wPBE-implementation and (M) imply the existence condition (b) of Definition 4. To prove the "only if" direction of Part (a), first, we show that if a mechanism Γ wr-implements a social choice function f, then Γ has a wPBE in every type space. This auxiliary result generalizes to dynamic mechanisms a result by BM from the static case, and simultaneously weakens the assumptions used in their result. Second, we note that if Γ wr-implements f then Γ satisfies (M).[23] Third, we obtain with minor modifications from Battigalli (1999) the implication that if Γ satisfies (M), then all wPBE strategies are weakly rationalizable. Once Part (a) is proven, Part (b) follows quickly from our work in Section 3.

Theorem 2 implies that under the assumption of the conditional NTI, excluding mechanisms that violate (M) from consideration is without of loss of generality, in the sense that this does not reduce the set of implementable social choice functions. As the statement of Theorem 2(a) confirms and as explained in Footnote 23, if Γ wr-implements f then Γ satisfies (M). Therefore, mechanisms that violate (M) do not wr-implement any social choice function and can be safely ignored. Parts (a) and (b) combined imply the analogous result for robust wPBE-implementation. If f is robustly wPBE-implementable then by part (b), f is wr-implementable and there is some Γ that wr-implements f. By Part (a), Γ satisfies (M) and robustly wPBE-implements f.

[23] If Γ implements f then $W^{\infty} \neq \emptyset$. Therefore, Γ satisfies (M) — if W_{-i}^{α} were not measurable then Π_i^{α} and therefore $W_i^{\alpha+1}$ would be empty by notational Convention (C), contradicting $W^{\infty} \neq \emptyset$.

On a more subtle note, the following example illustrates why condition (M) can nonetheless not be eliminated from Theorem 2(a), even if the conditional NTI is satisfied. While under the conditional NTI, a mechanism that violates (M) cannot expand the set of robustly wPBE-implementable social choice functions and cannot wr-implement any social choice function, it may still robustly wPBE-implement a social choice function. This constellation is only possible if the mechanism has uncountably many strategies for some player, even though the mechanism can be static and well-behaved (in the sense that each player has a best response at each information set against each belief).

Example 4.1 Let $I = 2$, $\Theta_i = \{\theta_i\}$ for all $i \in I$, $X = \{r_1, p_1\} \times \{r_2, p_2\}$, and

$$u_i(z_i) \stackrel{\text{def}}{=} u_i((z_1, z_2), \theta) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } z_i = r_i \\ 0 & \text{if } z_i = p_i \end{cases}.$$

This is a two player, complete information environment in which we can reward ($z_i = r_i$) or punish ($z_i = p_i$) each player i. Let E be non-measurable subset of the Euclidean space $\mathbb{R}$ and fix some $e \in E$. As we will see now, the non-measurability of E implies that the static mechanism which lets each player choose a real number and rewards i for choosing e or for matching s_{-i} as long as $s_{-i} \in E$ violates (M). For each i, let $S_i = \mathbb{R}$ be i's strategy set. For each $s \in S$, let the outcome assigned by the mechanism to s be[24]

$$C(s_1, s_2) = \begin{cases} (r_1, r_2) & \text{if } s_1 = s_2 \in E \\ (r_1, p_2) & \text{if } s_1 = e \neq s_2 \\ (p_1, r_2) & \text{if } s_1 \neq e = s_2 \\ (p_1, p_2) & \text{otherwise} \end{cases}.$$

Then for each i and $j \neq i$, $Q_i^1 = E$ and (M) is violated because the non-measurability of E implies that $W_{-j}^1 = W_i^1 = E \times \{\theta_i\}$ is non-measurable. Moreover, $\Pi_j^1 = \emptyset$ and $W_j^2 = \emptyset$ — since i's rational strategies comprise a non-measurable set, by the notational Convention (C) introduced in Subsection 4.2, $\mu_j(W_i^1|\{\emptyset\}) = 1$ is false for every CPS μ_j. Consequently, $W^\infty = \emptyset$. Nonetheless, a wPBE exists for all type spaces: Fix some $e' \in E$. For $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i \in \mathcal{I}}$, let $b_i(t_i) = e'$ and $g_i(t_i) = \delta(e')$ for all i, then $(b_i, g_i)_{i \in \mathcal{I}}$ is a wPBE. Moreover, by sequential rationality, if $(b_i, g_i)_{i \in \mathcal{I}}$ is a wPBE then $b_i(t_i) \in E$ for all i and $t_i \in T_i$. Thus, this mechanism robustly wPBE-implements f such that $f(\theta_1, \theta_2) = (r_1, r_2)$, even though it does not wr-implement it. Finally, this example satisfies the conditional NTI.

[24]Note that this mechanism satisfies the measurability condition of Definition 12, as for each i and s_i, $C(\zeta(s_i, \cdot)) : S_{-i} \to Y$ is measurable (for $s_1 = e$, $C(\zeta(s_1, \cdot))$ maps s_2 to (r_1, r_2) is $s_2 = e$ and to (r_1, p_2) if $s_2 \neq e$ and is thus obviously measurable; for $s_1 \in E \backslash \{e\}$, $C(\zeta(s_1, \cdot))$ maps s_2 to (r_1, r_2) if $s_2 = s_1$, to (p_1, r_2) if $s_2 = e$ and to (p_1, p_2) if $s_2 \notin \{e, s_1\}$; for $s_1 \notin E$, $C(\zeta(s_1, \cdot))$ maps s_2 to (p_1, r_2) is $s_2 = e$ and to (p_1, p_2) if $s_2 \neq e$).

4.4 Robust wPBE-Implementation in Countable Static Mechanisms

This subsection maintains the assumptions of this paper (including the finiteness of the payoff type spaces and the pure outcome space) and additionally imposes the following assumptions of BM, ensuring that the assumptions of both papers are met. First, this subsection focuses on countable static mechanisms. Second, this subsection restricts attention to countable type spaces when talking about robustness, that is, type spaces $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i\in\mathcal{I}}$ such that T_i is countable for all $i \in \mathcal{I}$. One can show that a countable mechanism Γ wr-implements a social choice function f if and only if Γ wPBE-implements f for all countable type spaces (see Müller, 2017a), rendering the latter restriction inconsequential.

BM, pioneering the analysis of full robust implementation, introduced a notion of rationalizable implementation in static mechanisms and showed that it is *almost* equivalent to robust implementation in interim equilibrium. Robust implementation in interim equilibrium ("robust implementation" in this subsection for brevity) is a static mechanism implementation concept that equals robust wPBE-implementation by static mechanisms.[25,26] The almost equivalence between rationalizable and robust implementation implies that the same social choice functions are implementable under both notions, at least if one only considers social choice functions that satisfy the conditional NTI (Definition 8). However, it can occur that a mechanism rationalizably but not robustly implements a social choice function, even if the conditional NTI is met.

Corollary 1 implies that wr-implementation by static mechanisms strengthens BM's definition of rationalizable implementation in such a way as to permit an *exact* equivalence: for every countable static mechanism Γ, wr-implementation is exactly equivalent to robust wPBE-implementation. This exact equivalence implies that the set of wr-implementable and robustly implementable social choice functions coincide, even if one permits social choice functions that violate the conditional NTI. Moreover, a designer who identifies a mechanism that wr-implements a social choice function is guaranteed that the mechanism also achieves robust implementation (and vice versa).

To formalize the discussion, let Γ be a countable static mechanism and f be a social choice function. The following Condition (A) equals Part (a) of our definition of wr-implementation

[25] For every static mechanism and every type space, the set of wPBE and the set of interim equilibria are identical for our purposes: If $(b_i, g_i)_{i\in\mathcal{I}}$ is a wPBE, then in the terminology of BM, $(b_i)_{i\in\mathcal{I}}$ is an interim (or Bayesian) equilibrium. Conversely, every profile $(b_i)_{i\in\mathcal{I}}$ of strategy functions uniquely determines an associated profile $(g_i)_{i\in\mathcal{I}}$ of belief functions via the consistency condition of Definition 10. In particular, if $(b_i)_{i\in\mathcal{I}}$ is an interim equilibrium, then $(b_i)_{i\in\mathcal{I}}$ together with the associated profile $(g_i)_{i\in\mathcal{I}}$ is a wPBE.

[26] By robust implementation, in accordance with Definitions 10 and 11, we mean robust implementation in *pure* interim equilibrium. While BM in fact study robust implementation in *mixed* interim equilibrium, their almost equivalence also applies to robust implementation in *pure* interim equilibrium (see Footnote 28). We further note that if Γ robustly implements f in pure interim equilibrium then Γ robust implements f in mixed interim equilibrium, but not vice versa (see Müller, 2017a). Hence, we focus on the more demanding of these two concepts here.

(Definition 4), and thus describes the implementation requirement that all weakly rationalizable strategy profiles lead to desired social outcomes. The following conditions (B1)-(B3) all imply that every payoff type has at least one weakly rationalizable strategy, but each strengthen this requirement to various degrees. Conditions (B1) and (B3) are from BM and Condition (B2) applies Definition 4(b) to static mechanisms.

(A) $C(s) = f(\theta)$ for all $(s, \theta) \in W^\infty$.

(B1) For each $i \in \mathcal{I}$ and $\psi_i \in \Delta(\Theta_{-i})$ there exists a $\lambda_i \in \Pi_i^\infty \subseteq \Delta(\Sigma_{-i})$ such that

- $\lambda_i(Q_{-i}^\infty(\theta_{-i}) \times \{\theta_{-i}\}) = \psi_i(\theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$ and
- $r_i(\theta_i, \lambda_i) \neq \emptyset$ for all $\theta_i \in \Theta_i$.

(B2) There exists a profile $(\mathcal{Q}_i(\theta_i))_{i\in\mathcal{I},\theta_i\in\Theta_i}$ of nonempty strategy sets $\mathcal{Q}_i(\theta_i) \subseteq Q_i^\infty(\theta_i)$ such that for all $i \in \mathcal{I}$, $\theta \in \Theta$ and $s_{-i} \in \mathcal{Q}_{-i}(\theta_{-i})$, there exists $s_i \in \mathcal{Q}_i(\theta_i)$ such that $s_i \in r_i(\theta_i, \delta(s_{-i}, \theta_{-i}))$.

(B3) (ex-post best response property for $(Q_i^\infty(\theta_i))_{i,\theta_i}$, BM) For all $i \in \mathcal{I}$ and $\theta_i \in \Theta_i$, there exists $s_i \in Q_i^\infty(\theta_i)$ such that $s_i \in r_i(\theta_i, \delta(s_{-i}, \theta_{-i}))$ for all $(s_{-i}, \theta_{-i}) \in W_{-i}^\infty$.

We obtain that

- Γ rationalizably implements (as defined by BM) f if and only if [(A) and (B1)],[27] that
- Γ wr-implements f if and only if [(A) and (B2)], and
- an additional notion of implementation introduced by BM, [(A) and (B3)].

These "rationalizability-based" implementation concepts are nested. In particular, for every countable static mechanism Γ and every social choice function f,

$$[(A) \text{ and } (B3)] \Rightarrow [(A) \text{ and } (B2)] \Rightarrow [(A) \text{ and } (B1)],$$

and the converse directions of these implications are false for some Γ and f (see Müller, 2017a).

BM, Theorem 3(2) show that [(A) and (B1)] is necessary for Γ to robustly implement f. However, [(A) and (B1)] is not sufficient for Γ to robust implement f. The key is that [(A) and (B1)] does not guarantee that Γ has an interim equilibrium on all type spaces (see Müller, 2017a). Consequently, BM resort to a stronger condition to establish sufficiency. In their Theorem 3(1), they prove that [(A) and (B3)] implies robust implementation. Together,

[27] See BM, Definition 4, noting that in static mechanisms, weak rationalizability reduces to rationalizability as defined by BM.

these two results establish an almost equivalence of rationalizable to robust implementation.[28] BM further show that it is impossible to achieve an exact equivalence by strengthening the conclusion of their Theorem 3(2) to [(A) and (B3)]. Corollary 1 achieves an exact equivalence between wr- and robust implementation by instead weakening the ex-post best response property (B3) to (B2). Property (B2) is implied by robust implementation, and at the same time suffices to establish the existence of an interim equilibrium for all type spaces.

In conclusion, wr-implementation by static mechanisms is the rationalizability-based implementation concept that exactly characterizes robust implementation. Wr-implementation by static mechanisms achieves this by embodying slightly weaker sufficient conditions than BM's [(A) and (B3)], and the slightly tighter necessary conditions than BM's [(A) and (B1)].

5 Discussion and Conclusion

We studied two implementation notions, robust wPBE-implementation and wr-implementation. Both notions are belief-free in that they assume no knowledge of the agents' initial belief hierarchies about the state of the world $\theta \in \Theta$, and belief-revision free in that they assume no knowledge of how agents revise their beliefs after encountering a surprise information set. We showed that these notions are equivalent (given a technical measurability condition), and introduced a dr-monotonicity condition that together with epIC is necessary, and under the conditional NTI condition, sufficient for robust wPBE- and wr-implementation. Our results apply to general dynamic mechanisms, but remain true if we restrict attention to countable mechanism.

5.1 Dynamic Mechanisms in Incomplete Information Settings in the Literature[29]

"Classical" Dynamic Implementation under Belief-Revision Assumptions. We now relate the current paper to Baliga (1999), who provides sufficient conditions for implementation in sequential equilibrium in economic environments, Brusco (1995, 1999), who examines implementation in Perfect Bayesian equilibrium (PBE), and Bergin and Sen (1998), who provide sufficient conditions for implementation in sequential equilibrium and related solution concepts.

Like us, all these papers consider incomplete information environments with finite payoff type spaces. Nonetheless, our paper is orthogonal to this important literature in several respects. First, while unlike these papers we restrict the outcome space to be a space of

[28] While BM state their Theorem 3 for robust implementation in *mixed* interim equilibrium, by their proof, their almost equivalence also applies to robust implementation in *pure* interim equilibrium, and thus the concept of robust implementation we focus on in this subsection (see Footnote 26).

[29] We skip a discussion of papers from the complete information literature such as Moore and Repullo (1988).

lotteries over pure outcomes, their sufficient conditions require at least three agents while ours also apply to the two agent case. Thus the set of environments to which our results apply overlaps but is different from the set of environment to which their results apply. Second, we examine implementation that does not depend on the initial belief hierarchies about payoff types, while the mentioned papers all assume a common prior. We thus primarily address designers that are concerned about the robustness of their mechanisms. Third, while this literature makes various belief-revision assumptions that go beyond Bayesian updating, a main feature of our analysis is that we forgo such assumptions.

Even neglecting the other differences, the third difference alone implies that our conditions for implementation and this literature's are not nested. Our analysis is based on wPBE, a weaker equilibrium concept than both PBE and sequential equilibrium. Since our sufficiency mechanism rules out all undesirable wPBE, it also rules out all undesirable PBE and sequential equilibria. Arguably, ruling out all undesirable equilibria is at the heart of full implementation. Even so, our results do not imply this literature's because our mechanism only guarantees the existence of a wPBE, and not the existence of a PBE or sequential equilibrium. Conversely, since ruling out all undesirable PBE and sequential equilibria does not imply ruling out all undesirable wPBE, our results do not follow from this literature's.

To expand on the third difference, it is fruitful to discuss some aspects of Bergin and Sen (1998). While these authors explicitly formulate their sufficient condition independent of the exact solution concept used, their sufficient condition does not apply to wPBE. This illustrates some salient differences between their and our assumptions (and clarifies that the mentioned independence extends only to sequential equilibrium and similar solution concepts).

Bergin and Sen (1998) focus on the possibility of generating preference reversals using posterior distributions. To illustrate their notion of a posterior distribution, suppose that it has been observed that some agent deviated from her (candidate) equilibrium strategy. Then the i-th component of a posterior distribution $\lambda \in \prod_{k \in \mathcal{I}} \Delta(\Theta_k)$ describes the beliefs that every $j \neq i$ holds about i's payoff type. Bergin and Sen (1998) formulate their sufficient condition, "posterior reversal," in terms of the set of posterior distributions permitted after such a deviation. The set of permitted posterior distributions induced by a particular solution concept may then, or not, satisfy the posterior reversal condition. The kind of beliefs permitted in a wPBE, however, are generally too rich to be captured by a posterior distribution. As indicated, using posterior distributions implicitly assumes that after a deviation, there is a common belief among all $j \neq i$ about i's payoff type. In wPBE, by contrast, every agent (except the agent that deviated) forms their own, individual, new belief after a deviation. Hence Bergin and Sen (1998)'s sufficient condition does not apply to wPBE, and our analysis complements theirs by encompassing richer sets of beliefs after surprises.

An additional indication that the posterior reversal condition is geared towards sequential

equilibrium and related concepts is that Bergin and Sen (1998)'s sufficiency proof implicitly assumes that even after a deviation from the equilibrium path, players are certain that their opponents will follow the equilibrium strategy "from now on." This is again not true in wPBE, which permits that after a surprise, players believe that their opponents no longer follow the equilibrium strategy. Finally, some instances of posterior reversals exploit that in sequential equilibrium a player i's beliefs about players j and k are independent, in the sense that a deviation by j does not influence i's belief about k (see e.g. Bergin and Sen, 1998, Example 2). Since wPBE allows e.g. that, after observing irrational play by j, player i places some probability on k also behaving irrationally in the future, such instances of posterior reversals would not exist given wPBE. Thus, while it already incorporates a variation of the idea of facilitating implementation by using preference reversals "off the equilibrium path," the posterior reversal condition is quite distinct from our robust preference reversal condition. In particular, our robust preference reversal condition does not assume any knowledge of the agents' belief revision.

In addition to their sufficient conditions, Bergin and Sen (1998) also provide necessary conditions for implementation. These, however, only apply to implementation by mechanisms "with one round of signaling," that is, mechanisms in which equilibrium play never goes beyond a first stage. It seems fair to say that the restriction to such mechanisms is not motivated by particular real-world concerns, but is imposed in order to keep the analysis tractable. In fact, Brusco (1999) shows that the restrictions to mechanisms with one round of signaling is with loss of generality by providing an example of an implementable social choice function that cannot be implemented by a mechanism with one round of signaling. Since Baliga (1999) only provides a sufficient but no necessary conditions for implementation, Brusco (1995, 2006) probably comes closest in this literature to a tight characterization of implementability of general dynamic mechanisms. Brusco (1995) allows general dynamic mechanisms, but there is a gap between his necessary and sufficient conditions. Brusco (2006) achieves a tight characterization, but restricts attention to multi-stage (and to avoid additional notational complexity, in fact, two-stage) mechanisms with public signals. Thus even in Brusco's (2006) case, one still may be concerned that one can enlarge the set of implementable social choice functions simply by removing an ad-hoc restriction on the class of admitted mechanisms.

Our approach does not face this difficulty. Our characterization applies to general dynamic mechanisms. In fact, the mechanism of our sufficiency result, Proposition 3, is neither multi-stage nor a mechanism with one round of signaling. Thanks to our simpler informational assumptions, and only apart from the mild conditional NTI property, we obtain a tight characterization of the set of implementable social choice functions. Formally complementing and not substituting for this literature's findings, from a broader perspective, our results may thus have some benchmark character for dynamic implementability in incomplete information

environments.

"Classical" Dynamic Implementation without Belief-Revision Assumptions. In an interesting contribution, Duggan (1998) constructs a set of outcomes from which each payoff type of agent i would pick one element as a strictly dominant choice. Only payoff types with identical utility functions would pick the same element, making this set of outcomes a "preference revelation device." Duggan (1998) then places such preference revelation devices off the equilibrium path in a two-stage mechanism in order to implement social choice functions. Notably, since his preference revelation devices give strictly dominant incentives, his implementation is independent of particular belief-revision assumptions.[30] Despite thus offering maximal robustness with respect to the belief-revision process, Duggan (1998) belongs to the classical implementation literature in that he assumes that agents have commonly known priors about the state of the world. In contrast, we adopt a belief-free approach. His and our implementation results are also complementary in other regards. In an informal sense, Duggan's mechanism is simpler than ours, but this comes at the cost of relying on more specialized assumptions. In particular, Duggan (1998) assumes quasi-linear preferences and rules out interdependent values by assuming private values, while our results apply to more general environments.[31] Duggan (1998) does not provide necessary conditions for implementation, while we characterize implementability.

Robust Dynamic Implementation. Finally, our paper also takes a different direction than Penta (2015) and Müller (2016), who already considered forms of robust implementation in dynamic mechanisms.

Penta (2015) provides sufficient conditions for robust implementation in multi-period environments in which an agent learns part of her payoff type in each period. Such environments are more general than the more classical static environments that we consider. Our analysis instead goes beyond Penta's (2015) in other ways. First, Penta (2015) restricts attention to mechanisms that are static each period and comprise a multi-stage mechanism with observable actions across periods. For his sufficient conditions, he indeed restricts attention to direct mechanisms. In our static environments, an agent learns her complete payoff type before the mechanism begins and later period payoff type revelations are trivial, making every direct mechanism a static mechanism. Static environments thus render Penta (2015)'s sufficient conditions ones for static implementation. One thus could say that Penta (2015) mostly focuses on dynamic *environments*, while we focus on dynamic *mechanisms*. Second, mostly advancing the

[30] He nonetheless adopts a different implementation concept than us. While we base our analysis on wPBE, he introduces a notion of double implementation that he terms implementation in sequentially rational strategies.

[31] Similar to us, Duggan (1998) also assumes a finite payoff type space and admits lotteries over a space as outcomes. Further assumptions of his include a no-total indifference condition and that the agents' priors have full support or at least common support.

theory of direct implementation, Penta (2015) does not characterize or even provide necessary conditions for full implementation by indirect mechanisms, one of our main contributions.

Further differences between Penta's (2015) and our work are that, first, Penta's (2015) sufficient conditions are for robust implementation in interim perfect equilibrium (IPE). IPE is stronger than wPBE, implying that Penta (2015) uses stronger belief-revision assumptions than we do here. Second, analogous to our Section 4, Penta (2015) provides an epistemic characterization that relates robust implementation in IPE to a novel notion he puts forward, backwards rationalizable implementation. However, since incentive compatibility and equilibrium existence coincide in direct mechanisms (see Penta, 2015, Proposition 1), for his purposes, a game theoretic equivalence (see beginning of our Section 4 and Footnote 15) suffices, while our epistemic characterization requires consideration of equilibrium existence in general mechanisms.

In Müller (2016), we provide necessary and sufficient conditions for (belief-free) robust implementation by finite dynamic mechanisms in static environments. The current paper also differs from this work in crucial aspects. First, in Müller (2016) we weaken the implementation concept to robust *virtual* implementation, and thus adopt an approximate notion of full implementation. In contrast, the current paper insists on robust *exact* implementation. Virtual implementation permits a vanishing but positive probability of implementing an outcome very different from the desired social outcome. Virtual implementation has its own advantages and disadvantages (see e.g. Abreu and Matsushima, 1992b,a; Glazer and Rosenthal, 1992) and, from a technical viewpoint, leads to a quite different analysis. Second, in Müller (2016) we examine strongly rationalizable implementation. We thus assume rationality and common strong belief in rationality (RCSBR) and impose stronger belief-revision assumptions than we do here, assuming that the agents engage in forward induction logic. This permits agents to "learn" their opponents' payoff types in an appropriately designed mechanism, which in turn facilitates the virtual implementation of incentive compatible social choice functions. This learning channel is at the core of the analysis in Müller (2016), but breaks down in the absence of forward induction. Third, the necessary and sufficient conditions in Müller (2016) coincide in "generic" private consumption environments that satisfy an economic property, but are not tight in general. Fourth, there is no known equilibrium concept that corresponds to strong rationalizability as wPBE corresponds to weak rationalizability. Hence the analysis in Müller (2016) is motivated directly by the appeal of RCSBR,[32] while our Section 4 provides an equilibrium based motivation for wr-implementation.

In a sense, Müller (2016) illustrates the potential of dynamic mechanisms for robust implementation by unveiling their advantage over static mechanisms in one scenario: in most private consumption environments, if the agents employ forward induction and one is content with

[32]See our Subsection 5.2 for a possible analogous motivation for wr-implementation.

virtual implementation, finite dynamic mechanisms can handle preference interdependencies much better than static mechanisms. In the current paper, we instead aim at a tight characterization of implementability for a pure form of robust exact implementation in a quite general setting. We believe such a characterization is useful as it reveals which social choice functions can be implemented without weakening the implementation concept, relying on stronger belief-revision assumptions, or relaxing wr-implementation in other ways. At the same time, such a characterization gives us certainty about the social choice functions that require some such relaxation to be implementable.

5.2 Weak Rationalizability and RCIBR

In this paper, our main motivation to study wr-implementation stemmed from its equivalence to robust wPBE-implementation. Since it does not involve the explicit formulation of type spaces, wr-implementation is easier to work with than and thus a useful proxy for robust wPBE-implementation. The value of our results on wr-implementation, however, likely extends further. The reason is that weak rationalizability characterizes the behavioral implications of the epistemic condition of rationality and common initial belief in rationality (RCIBR) (Battigalli and Siniscalchi, 2007).

Directly characterizing the set of social choice functions that are implementable for a given epistemic condition is an interesting complement to the more traditional, equilibrium based implementation theory. In this context, RCIBR is a baseline condition that (as expected) only assumes Bayesian updating with respect to the belief revision process. RCIBR captures that initially there is common belief in sequential rationality, but does not make assumptions about beliefs at surprise information sets. In Müller (2016), we used the more demanding condition of RCSBR, under which agents engage in forward induction reasoning after encountering a surprise.

While we do not pursue a formal analysis of implementation under RCIBR here, we expect our characterization of wr-implementation to be useful in such an endeavor. A technical difficulty that arises is that the epistemic relation between weak rationalizability and RCIBR has originally been established only for finite mechanisms, and thus a smaller class of mechanisms that we employ. While still not sufficient to immediately apply to our sufficiency mechanism, recently, however, Battigalli et al. (2017) extended this characterization to simple games, noting that they choose mechanisms with observable actions only for notational simplicity.

5.3 Conclusion

In conclusion, this paper provides general conditions for robust implementability in dynamic mechanisms, imposing Bayesian updating as the only belief-revision assumption. On the one hand, our conservative informational assumptions guarantee that all social choice functions

that satisfy our sufficient conditions for wr-implementation are implemented in a strong, robust way. In this context, the question of sufficient conditions for wr-implementation by well-behaved mechanisms is of future interest. On the other hand, the necessary conditions of our characterization describe restrictions to robust implementability. Since we already use a very general class of mechanisms, relaxing these necessary conditions will require stronger assumptions in other dimensions. In this context, our results can serve as a point of departure for future research in at least two directions. First, in some circumstances, mechanism designers may be comfortable to make stronger epistemic assumptions on the belief revision process. As discussed in Subsection 5.1, in Müller (2016), we already examined one specific instance of this approach for virtual implementation. Second, in other circumstances, stronger assumptions about initial beliefs may be acceptable. Ollár and Penta (2017) already provide interesting results in this direction for the static mechanism case.

A On the Definition of Weak Rationalizability

In Subsection 2.4 we defined weak rationalizability as W^∞ (Definition 3) instead of as W^{ω_0}. We thus require a strategy to survive more than ω_0 rounds of elimination of never-best sequential responses before we call it weakly rationalizable. In this appendix, we illustrate that the additional rounds of elimination required by W^∞ compared to W^{ω_0} are key to ensuring that wr-implementation is equivalent to robust wPBE-implementation. This appendix uses the notions of a type space, a wPBE and robust wPBE-implementation, see Definitions 9, 10 and 11 in Section 4.

Considering static games of complete information, Lipman (1994) shows that the iterated removal of never-best responses does not necessarily characterize common certainty of rationality if one uses only ω_0 rounds of elimination. The logic behind his result is our reason for insisting on more than ω_0 rounds of elimination in our definition of W^∞. The following example exhibits a countable mechanism that robustly wPBE-implements a social choice function f, but fails to wr-implement f if one uses W^{ω_0} as the definition of weak rationalizability. We were able to construct this example because in some mechanisms (indeed, even in some countable mechanisms), a fixed point is not yet reached after ω_0 rounds of elimination and, correspondingly, W^{ω_0} contains strategy-payoff type profiles that cannot arise in any equilibrium in any type space. Replacing W^{ω_0} with W^∞ renders robust wPBE- and wr-implementation equivalent, though, both in the example and also much more generally (see Theorem 2).

In the static mechanism case, BM implicitly also use transfinitely many rounds of elimination in their definition of (belief-free) rationalizability.

Example A.1 Let $I = 2$ and suppose that $\Theta_i = \{\theta_i\}$ for all $i \in \mathcal{I}$. That is, suppose that there is complete information among the two agents and the mechanism designer about the agents'

preferences. Let the set of pure outcomes be $X = \{-1, 0, 1\}^2$ and interpret $(x_1, x_2) \in X$ to mean that i receives x_i units of money. Accordingly, let $u_i(x_1, x_2) = x_i$ for all i and (x_1, x_2), where we suppress the trivial dependence of utility on θ from the notation (as we shall similarly do for other trivial dependencies).

Consider the social choice function f such that $f(\theta_1, \theta_2) = (0, 0)$. While the complete information assumed in this example makes it trivial to find *some* mechanism that robustly wPBE- and wr-implements f (not matter if we define weak rationalizability as W^{ω_0} or as W^{∞}), not *all* mechanisms that robustly wPBE-implement f also wr-implement f if we define weak rationalizability as W^{ω_0}. To see this, consider the static mechanism Γ with strategy set $S_i = \{-1, 0, 1, 2, \ldots\}$ for $i \in \mathcal{I}$ and the outcome function $C : S \to \Delta(X)$ such that

$$C(s) = \begin{cases} (1,0) & \text{if } s_1 > s_2 > 0 \\ (0,1) & \text{if } s_2 > s_1 > 0 \\ (1 - \frac{1}{2s_1}, 1) & \text{if } s_1 > s_2 = 0 \\ (1, 1 - \frac{1}{2s_2}) & \text{if } s_2 > s_1 = 0 \\ (0,0) & \text{if } s_1 = s_2 \\ (-1,-1) & \text{if } (s_1 = -1 \text{ or } s_2 = -1) \text{ and } s_1 \neq s_2 \end{cases},$$

where $(1 - \frac{1}{2s_1}, 1)$ denotes the lottery that places probability $(1 - \frac{1}{2s_1})$ on $(1, 1)$ and complementary probability $\frac{1}{2s_1}$ on $(0, 1)$, and $(1, 1 - \frac{1}{2s_2})$ denotes an analogous lottery.

Mechanism Γ robustly wPBE-implements f. First, for every type space $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i=1,2}$, the profile $(b_i, g_i)_{i=1,2}$ that for every i and $t_i \in T_i$ prescribes the strategy $b_i(t_i) = -1$ and belief $g_i(t_i) \overset{\text{def}}{=} g_i(t_i)(\cdot|\{\emptyset\}) = \delta(-1, \theta_{-i})$ is a wPBE. Therefore, for every type space, $(s_1, s_2) = (-1, -1)$ is realized in a wPBE. Second, for every type space, there is no wPBE (b_i, g_i) in which some type of some agents plays a strategy other than -1. Too see that, suppose that for $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)$ there is a wPBE (b_i, g_i) such that $A = \{a > 0 : \exists i \in \mathcal{I}, t_i \in T_i (a = b_i(t_i))\}$ is nonempty. Let $s^{\min} = \min A$. Let i be an agent and t_i be a type which plays $s^{\min}$ in equilibrium. Since $s_{-i} = 1, \ldots, s^{\min} - 1$ are not equilibrium strategies for any type of $-i$, we have $g_i(t_i)(\{-1, 0\} \cup \{s^{\min}, s^{\min} + 1, \ldots\}) = 1$. But then $s^{\min} \notin r_i(g_i(t_i))$, as either $s_i = -1$ or $s^{\min} + k$ for k large enough is a better response against $g_i(t_i)$. Contradiction. Next, suppose that there is a wPBE (b_i, g_i) such that $A = \emptyset$ but that $b_i(t_i) = 0$ for some i and t_i. Since $g_i(t_i)(\{-1, 0\}) = 1$, a contradiction arises because then $s_i = -1$ (if $g_i(t_i)(-1) = 1$) or $s_i = 1$ (if $g_i(t_i)(0) > 0$) is a better response than $b_i(t_i) = 0$.

Mechanism Γ does not wr-implement f if we define weak rationalizability as W^{ω_0}. Let $\lambda_i \in \Delta(S_j)$, where $i \in \mathcal{I}$ and $j \neq i$. If $\lambda_i(s_j = 0) > 0$ or if $\sup[supp(\lambda_i)] = \infty$, then for every $s_i \geq 0$ there is a $k \in \mathbb{N}$ such that $s_i + k$ is a better response than s_i against λ_i, and the set of best responses against λ_i is either empty (if $\lambda_i(s_j = -1)$ is sufficiently small) or comprises $s_i = -1$ (if $\lambda_i(s_j = -1)$ is sufficiently large). If $\lambda_i(s_j = 0) = 0$ and $\sup[supp(\lambda_i)] = m < \infty$,

then $s_i = -1$ (if $m = -1$ or more generally, if $\lambda_i(s_j = -1)$ is sufficiently large) or $s_i = 0$ and every $s_i > m$ (if $\lambda_i(s_j = -1)$ is sufficiently small [and by implication $m \geq 1$]) are best responses against λ_i. Therefore $Q_i^\alpha = \{-1, 0\} \cup \{\alpha + 1, \alpha + 2, \ldots\}$ for each $\alpha < \omega_0$, and thus $Q_i^{\omega_0} = \{-1, 0\}$. Hence Γ does not wr-implement f if we define weak rationalizability as W^{ω_0}.

Note that for Γ, W^{ω_0} does not form a fixed point of the iterated elimination procedure. However, if we iterate the process of deleting never-best sequential best responses one additional time, we arrive at the fixed point described by $Q_i^{\omega_0+1} = \{-1\}$ for all i. Therefore, for the transfinite definition of weak rationalizability, Γ wr-implements f and wr-implementation is equivalent to robust wPBE-implementation.

B Proofs

We immediately formulate all proofs in this appendix for general dynamic mechanisms as defined in Definition 12.

B.1 Proof of Proposition 1

Proof. As a preliminary step, note that for all $i \in \mathcal{I}$ and all $\theta_i, \theta_i' \in \Theta_i$, there does not exist a robust (θ_i', θ_i)-preference reversal if and only if

$$\begin{aligned}&\forall \psi_i' \in \Delta(\Theta_{-i}), y \in Y^{supp(\psi_i')} \exists \chi_i \in \Delta(supp(\psi_i') \times \Theta_{-i}) \forall x \in Y^{supp(\psi_i')} : \\ &E_{\psi_i'} u_i(x(\theta_{-i}), \theta_i', \theta_{-i}) \leq E_{\psi_i'} u_i(y(\theta_{-i}), \theta_i', \theta_{-i}) \Longrightarrow E_{\chi_i} u_i(x(\theta_{-i}'), \theta) \leq E_{\chi_i} u_i(y(\theta_{-i}'), \theta). \quad (10)\end{aligned}$$

We now show that if (10) holds, then for every mechanism $\Gamma = \langle A, H, (\mathbb{H}_i)_{i \in \mathcal{I}}, P, C \rangle$ and information set $\mathcal{H} \in \mathbb{H}_i$, if $s_i \in S_i(\mathcal{H})$ is a best response for θ_i' at $\mathcal{H}$ with respect to some belief $\lambda_i' \in \Delta(\Sigma_{-i}(\mathcal{H}))$, then there is a $\lambda_i \in \Delta(\Sigma_{-i}(\mathcal{H}))$ such that s_i is a best response for θ_i at $\mathcal{H}$ with respect to λ_i. Let $\psi_i' = marg_{\Theta_{-i}} \lambda_i'$. Moreover, for each $\theta_{-i} \in supp(\psi_i')$, define $(\lambda_i')^{\theta_{-i}} \in \Delta(S_{-i})$ by

$$(\lambda_i')^{\theta_{-i}}(B_{-i}) = \frac{\lambda_i'(B_{-i} \times \{\theta_{-i}\})}{\psi_i'(\theta_{-i})}$$

for all measurable $B_{-i} \subseteq S_{-i}$, and define $y \in Y^{supp(\psi_i')}$ by

$$y(\theta_{-i}) = \int_{S_{-i}(\mathcal{H})} C(\zeta(s)) d(\lambda_i')^{\theta_{-i}}(s_{-i})$$

for all $\theta_{-i} \in supp(\psi_i')$ (where the right-hand side is a Bochner integral). Note that

$$E_{\psi_i'} u_i(y(\theta_{-i}), \theta_i', \theta_{-i}) = \sum_{\theta_{-i}} \left[\int_{S_{-i}(\mathcal{H})} u_i \left(C(\zeta(s)), \theta_i', \theta_{-i} \right) d(\lambda_i')^{\theta_{-i}}(s_{-i}) \right] \psi_i'(\theta_{-i})$$

$$= \sum_{\theta_{-i}} \left[\sup \sum_k \left(\inf_{s_{-i} \in B^k_{-i}} u_i\big(C(\zeta(s)), \theta'_i, \theta_{-i}\big) \right) (\lambda'_i)^{\theta_{-i}}(B^k_{-i}) \right] \psi'_i(\theta_{-i})$$

$$= \sup \sum_k \left(\inf_{(s_{-i}, \theta_{-i}) \in A^k_{-i}} u_i\big(C(\zeta(s)), \theta'_i, \theta_{-i}\big) \right) \lambda'_i(A^k_{-i})$$

$$= E_{\lambda'_i} u_i(C(\zeta(s)), \theta'_i, \theta_{-i}), \tag{11}$$

where the first equality follows from the measurability of $C(\zeta(s_i, \cdot)) : S_{-i} \to Y$ and the linearity of u_i, and the suprema in the second and third lines extend over all finite partitions $\{B^k_{-i}\}$ of $S_{-i}(\mathcal{H})$ into measurable sets B^k_{-i} and all finite partitions $\{A^k_{-i}\}$ of $\Sigma_{-i}(\mathcal{H})$ into measurable sets A^k_{-i}, respectively. Let χ_i satisfy (10) for ψ'_i and y, and let

$$\lambda_i(B_{-i} \times \{\theta_{-i}\}) = \sum_{\theta'_{-i} \in supp(\psi'_i)} (\lambda'_i)^{\theta_{-i}}(B_{-i}) \chi_i(\theta'_{-i}, \theta_{-i})$$

for all measurable $B_{-i} \subseteq S_{-i}$ and all $\theta_{-i} \in \Theta_{-i}$. Note that

$$E_{\chi_i} u_i(y(\theta'_{-i}), \theta) = \sum_{(\theta'_{-i}, \theta_{-i})} \left(\int_{S_{-i}(\mathcal{H})} u_i\left(C(\zeta(s)), \theta\right) d(\lambda'_i)^{\theta'_{-i}}(s_{-i}) \right) \chi_i(\theta'_{-i}, \theta_{-i})$$

$$= \sum_{(\theta'_{-i}, \theta_{-i})} \left[\sup \sum_k \left(\inf_{s_{-i} \in B^k_{-i}} u_i\big(C(\zeta(s)), \theta\big) \right) (\lambda'_i)^{\theta'_{-i}}(B^k_{-i}) \right] \chi_i(\theta'_{-i}, \theta_{-i})$$

$$= \sum_{\theta_{-i}} \sup \sum_k \left(\inf_{s_{-i} \in B^k_{-i}} u_i\big(C(\zeta(s)), \theta\big) \right) \lambda_i(B^k_{-i} \times \{\theta_{-i}\})$$

$$= \sup \sum_k \left(\inf_{(s_{-i}, \theta_{-i}) \in A^k_{-i}} u_i\big(C(\zeta(s)), \theta\big) \right) \lambda_i(A^k_{-i})$$

$$= E_{\lambda_i} u_i(C(\zeta(s)), \theta), \tag{12}$$

where the suprema extend as above. Then for every $s'_i \in S_i(\mathcal{H})$, if we let

$$x(\theta_{-i}) = \int_{S_{-i}(\mathcal{H})} C(\zeta(s'_i, s_{-i})) d(\lambda'_i)^{\theta_{-i}}(s_{-i})$$

for all $\theta_{-i} \in supp(\psi'_i)$ and consider equations (11), (12) and analogous equalities for s'_i,

$$E_{\lambda'_i} u_i(C(\zeta(s'_i, s_{-i})), \theta'_i, \theta_{-i}) = E_{\psi'_i} u_i(x(\theta_{-i}), \theta'_i, \theta_{-i})$$

$$\leq E_{\psi'_i} u_i(y(\theta_{-i}), \theta'_i, \theta_{-i}) = E_{\lambda'_i} u_i(C(\zeta(s)), \theta'_i, \theta_{-i})$$

implies

$$E_{\lambda_i} u_i(C(\zeta(s'_i, s_{-i})), \theta) = E_{\chi_i} u_i(x(\theta'_{-i}), \theta) \leq E_{\chi_i} u_i(y(\theta'_{-i}), \theta) = E_{\lambda_i} u_i(C(\zeta(s)), \theta).$$

Suppose that mechanism Γ wr-implements f and that β is an unacceptable deception. Suppose by contradiction that β is not d-refutable. Then for all $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta'_i \in \beta_i(\theta_i)$, there does not exist a robust (θ'_i, θ_i)-preference reversal. Moreover, for each $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta'_i \in \beta_i(\theta_i)$, we can let $\theta_{-i}^{\theta'_i \leftarrow \theta_i} \in \Theta_{-i}$ and $\psi_i^{\theta'_i \leftarrow \theta_i} \in \Delta(\Theta_{-i})$ with $\psi_i^{\theta'_i \leftarrow \theta_i}(\beta_{-i}^{-1}(\theta_{-i}^{\theta'_i \leftarrow \theta_i})) = 1$ be such that

$$\forall x \in Y_i(\theta_{-i}^{\theta'_i \leftarrow \theta_i}) : \quad E_{\psi_i^{\theta'_i \leftarrow \theta_i}} u_i(x, \theta) \leq E_{\psi_i^{\theta'_i \leftarrow \theta_i}} u_i(f(\theta'_i, \theta_{-i}^{\theta'_i \leftarrow \theta_i}), \theta). \tag{13}$$

Below, we use a CPS with the initial belief that $-i$ "claims to be" payoff type $\theta_{-i}^{\theta'_i \leftarrow \theta_i}$ and that $-i$'s true payoff type is distributed according to $\psi_i^{\theta'_i \leftarrow \theta_i}$ in order to rationalize the lie θ'_i for θ_i.

For each $i \in \mathcal{I}$ and $\theta_i \in \Theta_i$, fix a $s_i(\theta_i) \in \mathcal{Q}_i(\theta_i)$ and a $\mu_i(\theta_i) \in \Pi_i^\infty$ such that $s_i(\theta_i)$ is a sequential best response for θ_i with respect to $\mu_i(\theta_i)$, and such that $\mu_i(\theta_i)(\cdot|\{\emptyset\}) = \delta(s_{-i}^{\theta_i, \{\emptyset\}}, \theta_{-i})$ for some $\theta_{-i} \in \Theta_{-i}$ and some $s_{-i}^{\theta_i, \{\emptyset\}} \in \mathcal{Q}_{-i}(\theta_{-i})$. Here, $\mathcal{Q}_i(\theta_i)$ is the set described in Definition 4. For convenience, let

$$N = \{(i, \theta_i, \theta'_i) : i \in \mathcal{I}, \theta_i, \theta'_i \in \Theta_i, \theta'_i \in \beta_i(\theta_i)\},$$

so that $(i, \theta_i, \theta'_i) \in N$ if θ'_i is an announcement of θ_i that β permits.

Step 1. Suppose that $(i, \bar{\theta}_i, \bar{\theta}'_i) \in N$. We claim that $s_i(\bar{\theta}'_i) \in Q_i^1(\bar{\theta}_i)$. Let μ_i be a CPS that satisfies the following conditions.

1. If we let $s_{-i}(\theta_{-i}^{\bar{\theta}'_i \leftarrow \bar{\theta}_i}) \stackrel{\text{def}}{=} \left(s_j(\theta_j^{\bar{\theta}'_i \leftarrow \bar{\theta}_i})\right)_{j \neq i}$, then $\mu_i((s_{-i}(\theta_{-i}^{\bar{\theta}'_i \leftarrow \bar{\theta}_i}), \theta_{-i})|\{\emptyset\}) = \psi_i^{\bar{\theta}'_i \leftarrow \bar{\theta}_i}(\theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$.

2. Suppose that $\mathcal{H} \in \mathbb{H}_i(s_i(\bar{\theta}'_i))$ is a surprise given the beliefs μ_i. Since $s_i(\bar{\theta}'_i)$ is a best response for $\bar{\theta}'_i$ at $\mathcal{H}$ against $\mu_i(\bar{\theta}'_i)(\cdot|\mathcal{H})$, and since there does not exist a robust $(\bar{\theta}'_i, \bar{\theta}_i)$-preference reversal, there is a $\lambda_i \in \Delta(\Sigma_{-i}(\mathcal{H}))$ such that $s_i(\bar{\theta}'_i)$ is a best response for $\bar{\theta}_i$ at $\mathcal{H}$ against λ_i. We require that $\mu_i(\cdot|\mathcal{H})$ equals such a λ_i.

By Condition 2., $s_i(\bar{\theta}'_i)$ maximises $\bar{\theta}_i$'s expected utility with respect to μ_i at all information sets that are admiited by $s_i(\bar{\theta}'_i)$ and that are or succeed a surprise information set (formally, at all $\mathcal{H} \in \mathbb{H}_i(s_i(\bar{\theta}'_i)) \backslash \mathbb{H}_i(s_{-i}(\theta_{-i}^{\bar{\theta}'_i \leftarrow \bar{\theta}_i}))$. Thus, to prove that $s_i(\bar{\theta}'_i) \in r_i(\bar{\theta}_i, \mu_i)$, we only need to verify that $s_i(\bar{\theta}'_i)$ maximizes $\bar{\theta}_i$'s expected utility with respect to $\mu_i(\cdot|\mathcal{H}')$ at all $\mathcal{H}' \in \mathbb{H}_i(s_i(\bar{\theta}'_i), s_{-i}(\theta_{-i}^{\bar{\theta}'_i \leftarrow \bar{\theta}_i}))$.

Pick an $x \in Y$. Suppose that $x \notin Y_i(\theta_{-i}^{\bar{\theta}'_i \leftarrow \bar{\theta}_i})$ and $C(\zeta(s_i, s_{-i}(\theta_{-i}^{\bar{\theta}'_i \leftarrow \bar{\theta}_i}))) = x$ for some $s_i \in S_i$.

Then

$$u_i(x, \theta_i'', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}) > u_i(f(\theta_i'', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}), \theta_i'', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})$$

for some $\theta_i'' \in \Theta_i$. Let $\mu_i' \in \Pi_i^\infty$ be such that 1) $\mu_i'(\cdot|\{\emptyset\})$ equals the degenerate belief in $(s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}), \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})$, and 2) there exists a sequential best response s_i^* for θ_i'' against μ_i'. Such a μ_i' exists by Definition 4. On the one hand, $C(\zeta(s_i^*, s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))) = f(\theta_i'', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})$ as $s_i^* \in Q_i^\infty(\theta_i'')$ and Γ wr-implements f. On the other hand, we must have $C(\zeta(s_i^*, s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))) \neq f(\theta_i'', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})$, as x provides i with more expected utility with respect to $\mu_i'(\cdot|\{\emptyset\})$ than $f(\theta_i'', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})$ and i believes that x is "in her reach." Contradiction. Consequently, if $x \in Y$ is such that $C(\zeta(s_i, s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))) = x$ for some $s_i \in S_i$, then $x \in Y_i(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})$ and, by (13),

$$E_{\psi_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}} u_i(x, \bar{\theta}_i, \theta_{-i}) \leq E_{\psi_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}} u_i(f(\bar{\theta}_i', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}), \bar{\theta}_i, \theta_{-i}).$$

Hence, at each $\mathcal{H} \in \mathbb{H}_i(s_i(\bar{\theta}_i'), s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))$, the strategy $s_i(\bar{\theta}_i')$ maximizes $\bar{\theta}_i$'s expected utility with respect to $\mu_i(\cdot|\mathcal{H})$ within $S_i(\mathcal{H})$.

Step 2. Let $\alpha \in Ord$. First, if α is a successor ordinal, then $s_i(\theta_i') \in Q_i^{\alpha-1}(\theta_i)$ for all $(i, \theta_i, \theta_i') \in N$ implies that $s_i(\theta_i') \in Q_i^\alpha(\theta_i)$ for all $(i, \theta_i, \theta_i') \in N$: For each $(i, \theta_i, \theta_i') \in N$, if μ_i denotes the CPS for constructed for (i, θ_i, θ_i') in Step 1, then $s_i(\theta_i') \in r_i(\theta_i, \mu_i)$. Moreover, since the support of μ_i is

$$\{s_{-i}(\theta_{-i}^{\theta_i' \leftarrow \theta_i})\} \times supp(\psi_i^{\theta_i' \leftarrow \theta_i}) \subseteq \{s_{-i}(\theta_{-i}^{\theta_i' \leftarrow \theta_i})\} \times \beta_{-i}^{-1}(\theta_{-i}^{\theta_i' \leftarrow \theta_i}),$$

μ_i is an element of $\Pi_i^{\alpha-1}$. Second, if α is a limit ordinal, then $s_i(\theta_i') \in Q_i^\beta(\theta_i)$ for all $(i, \theta_i, \theta_i') \in N$ and all $\beta < \alpha$ trivially implies that $s_i(\theta_i') \in Q_i^\alpha(\theta_i) = \bigcap_{\beta<\alpha} Q_i^\beta(\theta_i)$ for all $(i, \theta_i, \theta_i') \in N$.

Step 3. By Step 2, $s_i(\theta_i') \in Q_i^\infty(\theta_i)$ for all $(i, \theta_i, \theta_i') \in N$. As is easy to verify, since β is unacceptable, there are $i \in \mathcal{I}$, $\theta_i \in \Theta_i$, $\theta_i' \in \beta_i(\theta_i)$ and $\theta_{-i}^{\theta_i' \leftarrow \theta_i} \in \Theta_{-i}$ such that $f(\theta_i, \theta_{-i}^{\theta_i' \leftarrow \theta_i}) \neq f(\theta_i', \theta_{-i}^{\theta_i' \leftarrow \theta_i})$. Because Γ wr-implements f,

$$C(\zeta(s_i(\theta_i'), s_{-i}(\theta_{-i}^{\theta_i' \leftarrow \theta_i}))) = f(\theta_i, \theta_{-i}^{\theta_i' \leftarrow \theta_i}) \neq f(\theta_i', \theta_{-i}^{\theta_i' \leftarrow \theta_i}) = C(\zeta(s_i(\theta_i'), s_{-i}(\theta_{-i}^{\theta_i' \leftarrow \theta_i}))).$$

Contradiction. □

B.2 Proof of Proposition 2

Proof. We establish the claim by a direct proof. Suppose that Γ wr-implements f and, for each i and θ_i, let $\mathcal{Q}_i(\theta_i) \subseteq Q_i^\infty(\theta_i)$ be the set of strategies from Definition 4. Fix some $i \in \mathcal{I}$,

$\theta_i, \theta_i' \in \Theta_i$ and $\theta_{-i} \in \Theta_{-i}$. We are going to show that

$$u_i(f(\theta), \theta) \geq u_i(f(\theta_i', \theta_{-i}), \theta). \tag{14}$$

Pick some $s_{-i} \in \mathcal{Q}_{-i}(\theta_{-i})$. Then there exist a CPS $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ and a strategy $s_i \in \mathcal{Q}_i(\theta_i)$ such that $\mu_i(\cdot|\{\emptyset\})$ equals the degenerate belief in (s_{-i}, θ_{-i}) and such that $s_i \in r_i(\theta_i, \mu_i)$. In addition, there exists a $s_i' \in \mathcal{Q}_i(\theta_i')$. If $f(\theta_i', \theta_{-i}) = f(\theta)$ then (14) is trivially satisfied, thus consider the case that $f(\theta_i', \theta_{-i}) \neq f(\theta)$. By wr-implementation, $C(\zeta(s_i', s_{-i})) = f(\theta_i', \theta_{-i}) \neq f(\theta) = C(\zeta(s))$. So $\mathbb{H}_i(s) \neq \emptyset$ (otherwise $\zeta(s_i', s_{-i}) = \zeta(s)$). What is more, there is a (unique) information set $\mathcal{H}' \in \mathbb{H}_i(s)$ such that $s_i(\mathcal{H}') \neq s_i'(\mathcal{H}')$ and such that $s_i(\mathcal{H}) = s_i'(\mathcal{H})$ for all $\mathcal{H} \in \mathbb{H}_i(s)$ such that $\mathcal{H} \prec \mathcal{H}'$. By the definition of sequential rationality, $\forall \mathcal{H} \in \mathbb{H}_i(s_i) \forall \tilde{s}_i \in S_i(\mathcal{H}) : U_i^{\mu_i}(s_i, \theta_i, \mathcal{H}) \geq U_i^{\mu_i}(\tilde{s}_i, \theta_i, \mathcal{H})$. In particular,

$$U_i^{\mu_i}(s_i, \theta_i, \mathcal{H}') \geq U_i^{\mu_i}(s_i', \theta_i, \mathcal{H}'). \tag{15}$$

Since Γ wr-implements f and since $\mu_i((s_{-i}, \theta_{-i})|\mathcal{H}') = 1$,

$$\begin{aligned} U_i^{\mu_i}(s_i, \theta_i, \mathcal{H}') &= \int_{\Sigma_{-i}(\mathcal{H}')} u_i(C(\zeta(s)), \theta) d\mu_i((s_{-i}, \theta_{-i})|\mathcal{H}') \\ &= u_i(C(\zeta(s)), \theta) \\ &= u_i(f(\theta), \theta). \end{aligned}$$

Similarly, $U_i^{\mu_i}(s_i', \theta_i, \mathcal{H}') = u_i(f(\theta_i', \theta_{-i}), \theta)$ and (14) follows from (15). □

B.3 Proof of Theorem 2

For every mechanism, every type space $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i \in \mathcal{I}}$ and every $b = (b_i)_{i \in \mathcal{I}}$ with $b_i : T_i \to S_i$ for all i, let

$$\Sigma^b = \left\{ \left(b_i(t_i), \hat{\theta}_i(t_i)\right)_{i \in \mathcal{I}} \in \Sigma : t \in T \right\} \tag{16}$$

denote the set of strategy-payoff type profiles "realized" by b. The following lemma shows that for all mechanisms that satisfy (M), "every weakly perfect Bayesian equilibrium is weakly rationalizable." The lemma essentially equals Proposition 3.10 (1) of Battigalli (1999). We nonetheless include a proof, because some minor changes to Battigalli's proof are necessary in our set-up.[33]

Lemma 1 *For every mechanism Γ that satisfies (M), every type space $\mathcal{T}$ and every wPBE $(b_i, g_i)_{i \in \mathcal{I}}$, $\Sigma^b \subseteq W^\infty$.*

[33] We assume (M) but do not rely on Σ_{-i}^b being measurable. Also, we need to use transfinite induction.

Proof. Let $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i\in\mathcal{I}}$ be a type space and $(b_i, g_i)_{i\in\mathcal{I}}$ be a wPBE. Trivially, $\Sigma^b \subseteq W^0$ and for each i, $g_i(T_i) \subseteq \Pi_i^0$. Now suppose that for all $\alpha < \beta \in \text{Ord}$, both $\Sigma^b \subseteq W^\alpha$ and for all i, $g_i(T_i) \subseteq \Pi_i^\alpha$. If β is a limit ordinal, then $\Sigma^b \subseteq W^\beta = \bigcap_{\alpha<\beta} W^\alpha$. If β is a sucessor ordinal, then by the sequential rationality condition of wPBE,

$$\Sigma_i^b \subseteq \rho_i(g_i(T_i)) \subseteq \rho_i(\Pi_i^{\beta-1}) = W_i^\beta, \qquad \forall i \in \mathcal{I}.$$

Hence for all i, the set $\{t_{-i} \in T_{-i} : (b_{-i}(t_{-i}), \hat{\theta}_{-i}(t_{-i})) \in W_{-i}^\beta\}$ equals T_{-i} (and is therefore measurable), and the consistency condition of wPBE implies that for all $t_i \in T_i$,

$$g_i(t_i)(W_{-i}^\beta|\{\emptyset\}) = \tau_i(t_i)\{t_{-i} \in T_{-i} : (b_{-i}(t_{-i}), \hat{\theta}_{-i}(t_{-i})) \in W_{-i}^\beta\} = \tau_i(t_i)(T_{-i}) = 1$$

(recall that W_{-i}^β is measurable by (M)). Therefore, $g_i(T_i) \subseteq \Pi_i^\beta$ for all $i \in \mathcal{I}$. □

The next lemma shows that wr-implementation implies the existence of wPBEs in all type spaces. Thereby, it enhances in two directions the implication of BM, proof of Theorem 3(1) that rationalizable implementation by a static mechanism that satisfies the ex-post best response property (B3) implies the existence of a (pure strategy) interim equilibrium in all type spaces. First, as we alluded to in Subsection 4.4, Lemma 2 applied to static mechanisms uses weaker assumptions than BM, Theorem 3(1). Second, it generalizes their result to dynamic mechanisms.

Lemma 2 *If Γ wr-implements f then for every type space $\mathcal{T}$, Γ has a wPBE for $\mathcal{T}$.*

Proof. Since Γ wr-implements f, we can pick a nonempty $(\mathcal{Q}_i(\theta_i))_{i,\theta_i}$ such that for all $i \in \mathcal{I}$, $\theta \in \Theta$ and $s_{-i} \in \mathcal{Q}_{-i}(\theta_{-i})$, there exist $s_i'' \in \mathcal{Q}_i(\theta_i) \subseteq Q_i^\infty(\theta_i)$ and $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ such that $\mu_i((s_{-i}, \theta_{-i})|\{\emptyset\}) = 1$ and $s_i'' \in r_i(\theta_i, \mu_i)$ and thus

$$u_i(C(\zeta(s_i'', s_{-i})), \theta) \geq u_i(C(\zeta(s_i', s_{-i})), \theta) \quad \forall s_i' \in S_i \tag{17}$$

(where (17) is trivial if $\mathbb{H}_i(s_i'', s_{-i}) = \emptyset$ and follows from the optimality of s_i'' at the earliest information set in $\mathbb{H}_i(s_i'', s_{-i})$ otherwise). In fact, (17) remains true for all $i \in \mathcal{I}$, $\theta \in \Theta$ and $s_{-i} \in \mathcal{Q}_{-i}(\theta_{-i})$ if we replace s_i'' by any other $s_i \in \mathcal{Q}_i(\theta_i)$, as $u_i(C(\zeta(s_i'', s_{-i})), \theta) = u_i(f(\theta), \theta) = u_i(C(\zeta(s)), \theta)$. Therefore, for every $i \in \mathcal{I}$, $\theta \in \Theta$ and $s \in \mathcal{Q}(\theta) \stackrel{\text{def}}{=} \prod_{j\in\mathcal{I}} \mathcal{Q}_j(\theta_j)$,

$$u_i(C(\zeta(s)), \theta) \geq u_i(C(\zeta(s_i', s_{-i})), \theta) \quad \forall s_i' \in S_i. \tag{18}$$

For each i and θ_i, fix some arbitrary θ_{-i} and $s_{-i} \in \mathcal{Q}_{-i}(\theta_{-i})$ and let $s_i(\theta_i) \in \mathcal{Q}_i(\theta_i)$ and $\mu_i(\theta_i) \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ be such that $\mu_i(\theta_i)((s_{-i}, \theta_{-i})|\{\emptyset\}) = 1$ and $s_i(\theta_i) \in r_i(\theta_i, \mu_i(\theta_i))$. Then by (18), for each i and θ_i, $s_i(\theta_i)$ is a best response for θ_i in S_i with respect to every $\lambda_i \in \Delta(\Sigma_{-i})$

such that $\lambda_i\{(s'_{-i},\theta'_{-i}) \in \Sigma_{-i} : s'_{-i} = s_{-i}(\theta'_{-i})\} = 1$. Suppose that $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i\in\mathcal{I}}$ is a type space. Define $(b_i, g_i)_{i\in\mathcal{I}}$ by letting, for all i, $b_i : T_i \to S_i$ be such that $b_i(t_i) = s_i(\hat{\theta}_i(t_i))$ for all t_i, and $g_i : T_i \to \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ be such that $g_i(t_i)$ is the CPS such that

- $g_i(t_i)(B_{-i}|\{\emptyset\}) = \tau_i(t_i)\{t_{-i} \in T_{-i} : (b_{-i}(t_{-i}), \hat{\theta}_{-i}(t_{-i})) \in B_{-i}\}$ for all measurable B_{-i} (note that $\{t_{-i} \in T_{-i} : (b_{-i}(t_{-i}), \hat{\theta}_{-i}(t_{-i})) \in B_{-i}\}$ is measurable as it equals the finite union $\cup_{\theta_{-i}\in\Theta_{-i}:(s_{-i}(\theta_{-i}),\theta_{-i})\in B_{-i}} \hat{\theta}_{-i}^{-1}(\theta_{-i})$ of measurable sets $\hat{\theta}_{-i}^{-1}(\theta_{-i}) = \prod_{j\neq i} \hat{\theta}_j^{-1}(\theta_j)$; further note that $g_i(t_i)(\{(s'_{-i},\theta'_{-i}) \in \Sigma_{-i} : s'_{-i} = s_{-i}(\theta'_{-i})\}|\{\emptyset\}) = 1$, and thus that $s_i(\theta_i)$ is a best response against $g_i(t_i)(\cdot|\{\emptyset\}))$, and
- if $\mathcal{H} \in \bar{\mathbb{H}}_i$ is a surprise, then $g_i(t_i)(\cdot|\mathcal{H}) = \mu_i(\hat{\theta}_i(t_i))(\cdot|\mathcal{H})$,

for all t_i. Then $(b_i, g_i)_{i\in\mathcal{I}}$ is a wPBE, because for every i,

- b_i is measurable, as for every $S'_i \subseteq S_i$, $b_i^{-1}(S'_i)$ equals the finite union $\cup_{\theta_i\in\Theta_i:s_i(\theta_i)\in S'_i} \hat{\theta}_i^{-1}(\theta_i)$ of measurable sets $\hat{\theta}_i^{-1}(\theta_i)$, and
- by construction, $b_i(t_i)$ is a sequential best response for $\hat{\theta}_i(t_i)$ against $g_i(t_i)$, for all $t_i \in T_i$, and
- by construction, the consistency condition of a wPBE is satisfied, for all $t_i \in T_i$.

□

To prove the following lemma, we show that W^∞ corresponds to a fixed point of a sequential best response operator, and thus to a wPBE. The underlying idea of this approach is familiar from related results in the literature. What somewhat complicates our case is that we prove the lemma for a general class of mechanisms. For example, Brandenburger and Dekel (1987) focus on finite games for which the fixed point property of rationalizability is more immediate, and depart from a best-reply sets definition of rationalizability from the start. For their respective solution concepts, Bernheim (1984) and Battigalli (1999) exploit compactness and continuity properties that we do not assume here. In BM, Tarski's fixed point theorem applies, while in our case, a non-monotonicity of the sequential best response operator prevents a direct application of Tarski's theorem. We solve this by adapting arguments from Echenique (2005).

Lemma 3 *If Γ robustly wPBE-implements f, then $C(\zeta(s)) = f(\theta)$ for all $(s,\theta) \in W^\infty$.*

Proof. First, we represent W^∞ as a fixed point of an appropriate operator. For all elements $\mathcal{S}$ and $\mathcal{S}'$ of the set $\mathscr{S} \stackrel{\text{def}}{=} \left\{\mathcal{S} = (\mathcal{S}_i(\theta_i))_{i,\theta_i} : \mathcal{S}_i(\theta_i) \subseteq S_i \text{ for all } i \in \mathcal{I} \text{ and } \theta_i \in \Theta_i\right\}$, write $\mathcal{S} \leq \mathcal{S}'$ if and only if $\mathcal{S}_i(\theta_i) \subseteq \mathcal{S}'_i(\theta_i)$ for all $i \in \mathcal{I}$ and $\theta_i \in \Theta_i$. One element of $\mathscr{S}$ is $\mathcal{W} = (Q_i^\infty(\theta_i))_{i,\theta_i}$. Define an operator $b = (b_1, \ldots, b_I) : \mathscr{S} \to \mathscr{S}$ by

$$b_i(\mathcal{S})(\theta_i) = \left\{ s_i \in S_i : \exists \mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i}) \right.$$

$$\left(\mu_i\left(\bigcup_{\theta_{-i}\in\Theta_{-i}}\left(\mathcal{S}_{-i}(\theta_{-i})\times\{\theta_{-i}\}\right)\middle|\{\emptyset\}\right)=1 \text{ and } (s_i,\theta_i)\in\rho_i(\mu_i)\right)\Bigg\}$$

While $(\mathscr{S},\leq)$ is a complete lattice, b can be non-monotone.[34] We thus can generally not apply Tarksi's fixed point theorem to b. Fortunately, we can still derive that $\mathcal{W}$ is a fixed point of b by following and adapting some steps of Echenique (2005, Lemma 1) (see also Echenique's references regarding his Lemma 1). Let γ be an ordinal number with cardinality greater than that of $\mathscr{S}$. Define $f:(\gamma+1)\to\mathscr{S}$ by transfinite recursion by $[f(0)]_i(\theta_i)=S_i$ and $[f(\beta)]_i(\theta_i)=\bigcap_{\alpha<\beta}b_i(f(\alpha))(\theta_i)$ for $0<\beta<\gamma+1$, for all i and θ_i. Even though b can be non-monotone, f is weakly decreasing by definition, that is, $\alpha<\beta$ implies $f(\beta)\leq f(\alpha)$. Thus for all $\alpha<\gamma$, $f(\alpha+1)=b(f(\alpha))$, and $[f(\alpha)]_i(\theta_i)=Q_i^\alpha(\theta_i)$ for all $\alpha<\gamma+1$, i and θ_i. Since the cardinality of γ is greater than the cardinality of $\mathscr{S}$, there is an $\alpha<\gamma$ such that $f(\alpha)=f(\alpha+1)$. Let $\bar{\alpha}$ be the smallest such α, then $f(\bar{\alpha})=f(\bar{\alpha}+1)=b(f(\bar{\alpha}))$ and $f(\bar{\alpha})$ is a fixed point of b. Finally, $f(\bar{\alpha})=\mathcal{W}$.

Second, since the claim of the lemma is trivial if W^∞ is empty, assume $W^\infty\neq\emptyset$. Since $\mathcal{W}$ is a fixed point of b, for each $(s_i,\theta_i)\in W_i^\infty$ there is a $g_i(s_i,\theta_i)\in\Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ such that $g_i(s_i,\theta_i)(W_{-i}^\infty|\{\emptyset\})=1$ and $s_i\in r_i(\theta_i,g_i(s_i,\theta_i))$. For each $i\in\mathcal{I}$, let $T_i=W_i^\infty$ and endow it with the relative topology inherited from Σ_i. Let $\hat{\theta}_i$ be the projection from T_i to Θ_i and define $\tau_i:T_i\to\Delta(T_{-i})$ by letting $\tau_i(s_i,\theta_i)$ equal the restriction of $g_i(s_i,\theta_i)(\cdot|\{\emptyset\})$ to the Borel σ-algebra on T_{-i}, for all $(s_i,\theta_i)\in T_i$. Then $(T_i,\hat{\theta}_i,\tau_i)_{i\in\mathcal{I}}$ is a type space because $\hat{\theta}_i$ is continuous and thus measurable. Moreover, if b_i denotes the projection from T_i to S_i, then b_i is measurable, and $(b_i,g_i)_{i\in\mathcal{I}}$ (with $g_i:T_i\to\Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ as just defined) satisfies by construction the sequential rationality and consistency conditions of a wPBE for $\mathcal{T}$. Since $(b_i,g_i)_{i\in\mathcal{I}}$ is a wPBE for $\mathcal{T}$, the lemma follows from the definition of robust wPBE-implementation. □

Finally, we prove the following lemma.

Lemma 4 *If Γ robustly wPBE-implements f, then there exists a profile $(\mathcal{Q}_i(\theta_i))_{i\in\mathcal{I},\theta_i\in\Theta_i}$ of nonempty strategy sets such that for all $i\in\mathcal{I}$, $\theta\in\Theta$ and $s_{-i}\in\mathcal{Q}_{-i}(\theta_{-i})$, there exist $s_i\in\mathcal{Q}_i(\theta_i)$ and $\mu_i\in\Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ such that $\mu_i((s_{-i},\theta_{-i})|\{\emptyset\})=1$ and $s_i\in r_i(\theta_i,\mu_i)$. If Γ also satisfies (M), then in addition $\mathcal{Q}_i(\theta_i)\subseteq Q_i^\infty(\theta_i)$ for all $i\in\mathcal{I}$ and $\theta\in\Theta$.*

Proof. We start by defining a type space that captures all coherent belief hierarchies comprised of only degenerate beliefs. For each $i\in\mathcal{I}$, let $T_i^0=\Theta_i$. Recursively define $T_i^k=\prod_{j\neq i}T_j^{k-1}$, for all $i\in\mathcal{I}$ and all $k\in\{1,2,3,\ldots\}$. For all i and k, endow T_i^k with the discrete

[34] If $\mathcal{S}\leq\mathcal{S}'$ and $\bigcup_{\theta_{-i}\in\Theta_{-i}}\left(\mathcal{S}'_{-i}(\theta_{-i})\times\{\theta_{-i}\}\right)$ is measurable for all i, then $b(\mathcal{S})\leq b(\mathcal{S}')$; however, if $\mathcal{S}\leq\mathcal{S}'$, $b(\mathcal{S})\neq\emptyset$ and $\bigcup_{\theta_{-i}\in\Theta_{-i}}\left(\mathcal{S}'_{-i}(\theta_{-i})\times\{\theta_{-i}\}\right)$ is non-measurable for some i, then, according to the notational Convention (C) from Subsection 4.2, $\emptyset\neq b(\mathcal{S})\not\leq b(\mathcal{S}')=\emptyset$.

topology. For each i, let $T_i = \prod_{k\in\mathbb{N}} T_i^k$ be the set of i's types and endow it with the product topology. Let $\hat{\theta}_i$ be the projection from T_i to Θ_i, that is, let

$$\hat{\theta}_i(t_i^0, t_i^1, t_i^2, \ldots) = t_i^0, \qquad \text{for all } (t_i^0, t_i^1, t_i^2, \ldots) \in T_i.$$

Finally, define $\tau_i : T_i \to \Delta(T_{-i})$ by

$$\tau_i(t_i^0, (t_{ij}^1)_{j\neq i}, (t_{ij}^2)_{j\neq i}, \ldots) = \delta((t_{ij}^1, t_{ij}^2, t_{ij}^3, \ldots)_{j\neq i}), \qquad \text{for all } (t_i^0, (t_{ij}^1)_{j\neq i}, (t_{ij}^2)_{j\neq i}, \ldots) \in T_i,$$

where $t_{ij}^k \in T_j^{k-1}$ for all $j \neq i$ and $k \geq 1$. Then $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i\in\mathcal{I}}$ is a type space, since the maps $\hat{\theta}_i$ are continuous by definition and thus measurable. Type $(t_i^0, t_i^1, t_i^2, \ldots)$ has payoff type t_i^0, is certain that his opponents' payoff types are $t_i^1 = (t_{ij}^1)_{j\neq i} \in T_i^1 = \prod_{j\neq i} \Theta_j$, and that $j \neq i$ believes that j's opponents' payoff types are t_{ij}^2 and so on.

Because Γ robustly wPBE-implements f, there exists a wPBE $(b_i, g_i)_{i\in\mathcal{I}}$ for $\mathcal{T}$. Let Σ^b be defined as in (16), and let $\mathcal{Q}_i(\theta_i) = \{s_i \in S_i : (s_i, \theta_i) \in \Sigma_i^b\}$ be the section of Σ_i^b at θ_i. For every $i \in \mathcal{I}$, $\theta \in \Theta$ and $s_{-i} \in \mathcal{Q}_{-i}(\theta_{-i})$, let, for each $j \neq i$, $(t_j^0, t_j^1, t_j^2, \ldots) \in b_j^{-1}(s_j)$ be such that $t_j^0 = \theta_j$, and $s_i = b_i(\theta_i, (t_j^0)_{j\neq i}, (t_j^1)_{j\neq i}, \ldots)$. Then $s_i \in \mathcal{Q}_i(\theta_i)$ and there exists $\mu_i \in \Delta^{\bar{\mathbb{H}}_i}(\Sigma_{-i})$ — namely, $g_i(t_i)$ for $t_i = (\theta_i, (t_j^0)_{j\neq i}, (t_j^1)_{j\neq i}, \ldots) \in T_i$ — such that

$$\mu_i((s_{-i}, \theta_{-i})|\{\emptyset\}) = \tau_i(t_i)\{t_{-i} \in T_{-i} : (b_{-i}(t_{-i}), \hat{\theta}_{-i}(t_{-i})) = (s_{-i}, \theta_{-i})\} = 1$$

(by the consistency condition of $(b_i, g_i)_{i\in\mathcal{I}}$ for t_i and since $(t_j^0, t_j^1, t_j^2, \ldots)_{j\neq i} \in \{t_{-i} \in T_{-i} : (b_{-i}(t_{-i}), \hat{\theta}_{-i}(t_{-i})) = (s_{-i}, \theta_{-i})\}$) and $s_i \in r_i(\theta_i, \mu_i)$ (by the sequential rationality condition of $(b_i, g_i)_{i\in\mathcal{I}}$ for t_i).

Finally, if Γ satisfies (M), Lemma 1 implies $\Sigma^b \subseteq W^\infty$ and thus $\mathcal{Q}_i(\theta_i) \subseteq Q_i^\infty(\theta_i)$ for all i and θ_i. □

If Γ wr-implements f then Γ satisfies (M) (see Footnote 23). Thus Lemmata 1 and 2 imply that if Γ wr-implements f, then Γ satisfies (M) and robustly wPBE-implements f. Lemmata 3 and 4 imply the converse implication. This proves part (a) of Theorem 2. Part (b) follows from part (a) and Proposition 3 if one shows that robust wPBE-implementation by all mechanisms, even those that violate (M), implies dr-monotonicity and epIC. Since this proof is analogous to the proofs of Propositions 1 and 2, we do not detail it here but just make the following remarks.

The first sentence of Lemma 4 still applies if we do not assume (M), even though $\mathcal{Q}_i(\theta_i) \subseteq Q_i^\infty(\theta_i)$ is no longer guaranteed. Moreover, in that first sentence, $\mathcal{Q}_i(\theta_i)$ is the set of θ_i's equilibrium strategies and the CPSs μ_i are equilibrium beliefs of the wPBE of the type space $\mathcal{T} = (T_i, \hat{\theta}_i, \tau_i)_{i\in\mathcal{I}}$ introduced in the proof of Lemma 4. Essentially, for proving that robust

wPBE-implementation implies dr-monotonicity and epIC, the $\mathcal{Q}_i(\theta_i)$ from Lemma 4 replaces the $\mathcal{Q}_i(\theta_i)$ from Definition 4 in the proofs of Propositions 1 and 2.

To adapt the proof of Proposition 1, note that it is easy to extend $\mathcal{T}$ from the proof of Lemma 4 (in fact, to extend every given type space) by a finite number of types per agent. In extending $\mathcal{T}$ to $\mathcal{T}'$, we can choose the type function of i to be every arbitrary extension of $\hat{\theta}_i$ and the belief function of i to be every arbitrary extension of τ_i. If (b_i, g_i) is a wPBE of $\mathcal{T}$, then (b'_i, g'_i) is a wPBE of $\mathcal{T}'$ if b'_i is an extension of b_i, g'_i is an extension of g_i, $g_i(t_i)$ is defined by the consistency condition of wPBE for all added types t_i, and $b_i(t_i)$ is sequentially rational for $\hat{\theta}'_i(t_i)$ with respect to $g_i(t_i)$ for all added types t_i.

The proof of Proposition 1 shows that for each $(i, \theta_i, \theta'_i) \in N$, $s_i(\theta'_i)$ is weakly rationalizable for θ_i. In the adapted proof, we can extend $\mathcal{T}$ from the proof of Lemma 4 by one type for each element of N. We can choose the extension so that the equilibrium (b_i, g_i) of $\mathcal{T}$ extends to a wPBE of the extended type space. Specifically, we can let the type corresponding to (i, θ_i, θ'_i) be a type of player i, have payoff type θ_i and $s_i(\theta'_i)$ as equilibrium strategy.

Adapting the proof of Proposition 2 requires only minor changes.

References

Abreu, D. and H. Matsushima (1992a): "A Response to Glazer and Rosenthal," *Econometrica*, 60, 1439–1442.

——— (1992b): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993–1008.

Abreu, D. and A. Sen (1990): "Subgame Perfect Implementation: A Necessary and Almost Sufficient Condittion," *Journal of Economic Theory*, 50, 285–299.

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

Baliga, S. (1999): "Implementation in Economic Environments with Incomplete Information: The Use of Multi-Stage Games," *Games and Economic Behavior*, 27, 173–183.

Battigalli, P. (1999): "Rationalizability in Incomplete Information Games," Working Paper.

——— (2003): "Rationalizability in infinite, dynamic games with incomplete information," *Research in Economics*, 57, 1–38.

Battigalli, P., G. Beneduci, and P. Tebaldi (2017): "Interactive Epistemology in Simple Dynamic Games with a Continuum of Strategies," Working Paper.

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

——— (2007): "Interactive epistemology in games with payoff uncertainty," *Research in Economics*, 61, 165–184.

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

——— (2009a): "Robust Implementation in Direct Mechanisms," *Review of Economic Studies*, 76, 1175–1204.

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

——— (2011): "Robust implementation in general mechanisms," *Games and Economic Behavior*, 71, 261–281.

Bergin, J. and A. Sen (1998): "Extensive Form Implementation in Incomplete Information Environments," *Journal of Economic Theory*, 80, 222–256.

Bernheim, B. D. (1984): "Rationalizable Strategic Behavior," *Econometrica*, 52, 1007–1028.

Bogachev, V. I. (2007): *Measure Theory, Volume 2*, Springer-Verlag.

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

Brusco, S. (1995): "Perfect Bayesian implementation," *Economic Theory*, 5, 419–444.

——— (1999): "Implementation with Extensive Form Games: One Round of Signaling Is Not Enough," *Journal of Economic Theory*, 87, 356–378.

——— (2006): "Perfect Bayesian implementation in economic environments," *Journal of Economic Theory*, 129, 1–30.

Chung, K.-S. and J. C. Ely (2007): "Foundations of Dominant-Strategy Mechanisms," *Review of Economic Studies*, 74, 447–476.

Duggan, J. (1998): "An extensive form solution to the adverse selection problem in principal/multi-agent environments," *Review of Economic Design*, 3, 167–191.

Echenique, F. (2005): "A Short and Constructive Proof of Tarski's Fixed-Point Theorem," *International Journal of Game Theory*, 33, 215–218.

Glazer, J. and R. W. Rosenthal (1992): "A Note on Abreu-Matsushima Mechanisms," *Econometrica*, 60, 1435–1438.

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

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

Moore, J. and R. Repullo (1988): "Subgame Perfect Implementation," *Econometrica*, 56, 1191–1220.

Müller, C. (2016): "Robust Virtual Implementation under Common Strong Belief in Rationality," *Journal of Economic Theory*, 162, 407–450.

——— (2017a): "A Note on Equilibrium Existence in Robust Implementation," Work in Progress.

——— (2017b): "On Weakly Rationalizable Implementation," Work in Progress.

Ollár, M. and A. Penta (2017): "Full Implementation and Belief Restrictions," *American Economic Review*, 107, 2243–2277.

Osborne, M. J. and A. Rubinstein (1994): *A Course in Game Theory*, The MIT Press.

Penta, A. (2012): "Higher Order Uncertainty and Information: Static and Dynamic Games," *Econometrica*, 80, 631–660.

——— (2013): "On the structure of rationalizability for arbitrary spaces of uncertainty," *Theoretical Economics*, 8, 405–430.

——— (2015): "Robust Dynamic Implementation," *Journal of Economic Theory*, 160, 280–316.

Rényi, A. (1955): "On a new axiomatic theory of probability," *Acta Mathematica Academiae Scientiarum Hungaricae*, 6, 285–335.

Rubinstein, A. (1989): "The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge"," *American Economic Review*, 79, 385–391.

Vartiainen, H. (2007): "Subgame perfect implementation: A full characterization," *Journal of Economic Theory*, 133, 111–126.

Weinstein, J. and M. Yildiz (2007): "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements," *Econometrica*, 75, 365–400.