Robust Implementation in Weakly Rationalizable Strategies*

Christoph Müller†
Carnegie Mellon University

February 12, 2013

Abstract

Weakly rationalizable implementation represents a generalization of robust implementation to dynamic mechanisms. It is so conservative that virtual implementation in weakly rationalizable strategies is characterized by the same conditions as robust virtual implementation by static mechanisms. We show that despite that, (exact) weakly rationalizable implementation is more permissive than (exact) robust implementation in general static mechanisms. We introduce a dynamic robust monotonicity condition that is weaker than Bergemann and Morris' (2011) robust monotonicity condition and prove that it is necessary, and together with weak extra assumptions sufficient for weakly rationalizable implementation in general dynamic mechanisms. We demonstrate that sometimes even weakly rationalizable implementation in finite dynamic mechanisms is more permissive than robust implementation in general static mechanisms.

KEYWORDS: robust implementation, weak rationalizability, common initial belief in rationality, dynamic robust monotonicity, strategic distinguishability.

*This paper originated in part of my Ph.D. thesis, Müller (2010). I thank my advisor Kim-Sau Chung for his encouragement and guidance, David Rahman for many inspiring conversations, Itai Sher and Jan Werner. Financial support through a Doctoral Dissertation Fellowship of the Graduate School of the University of Minnesota is gratefully acknowledged. All errors are my own.

†Contact: cmueller@cmu.edu

1 Introduction

Equilibrium outcomes of games and mechanisms are sensitive even to minor changes in the agents' beliefs and higher order beliefs about each others' payoff types. Yet basically any typical Bayesian model makes strong, implicit common knowledge assumptions about its agents' beliefs and higher order beliefs. This motivates the search for mechanisms that do not depend on such "details." Following the so-called Wilson doctrine (Wilson, 1987), the desire is to weaken or eliminate strong common knowledge assumptions and derive mechanisms that are more robust. Chung and Ely (2007) demonstrate that the desire for "detail-free" mechanisms can justify the use of dominant-strategy mechanisms. Bergemann and Morris (2005, 2009a,b, 2011) introduce and examine robust implementation in static mechanisms. In this paper, we study a strong form of robust implementation in dynamic mechanisms, namely (full) implementation in weakly rationalizable strategies.[1]

Many models in applied game theory use Bayesian Nash equilibrium to solve static games, and implicitly admit only one possible belief (and higher order belief) per payoff type. To arrive at robust static mechanisms, Bergemann and Morris explicitly break this one-to-one relation and allow any belief to be associated with any payoff type. In Bergemann and Morris (2011) they show that (full) robust static implementation essentially amounts to (full) rationalizable implementation. Analogous arguments motivate weakly rationalizable implementation as a robust form of implementation in weak perfect Bayesian equilibrium. We describe this approach in more detail later in this introduction on page 4.

Prior to that we want to put weakly rationalizable implementation into context relative to other, existing notions of robust implementation for dynamic mechanisms. For this purpose, it is useful to adopt a different angle and gauge a mechanism directly by the epistemic assumption on which it relies. The key feature a robust mechanism should possess is that it should not rely on any assumptions about the beliefs about others' payoff types. Static mechanisms that rationalizably implement a social choice function satisfy this criterion because rationalizability is characterized by rationality and common belief in rationality (RCBR) (Tan and Werlang, 1988). RCBR says that every agent is rational, believes everybody is rational, believes that everybody believes that everybody is rational and so on, and so does clearly not impose any implicit assumptions about the agents' beliefs and higher order beliefs about others' payoff types. Importantly, neither do the generalizations of RCBR to dynamic mechanisms that we describe below and that correspond to several distinct notions of robustness for dynamic mechanisms.

In a dynamic mechanism, agents have multiple beliefs, one at each information set. These beliefs are Bayesian updates of each other whenever possible; however, if an agent is surprised

[1] Full implementation requires that *every* weakly rationalizable outcome coincides with the outcome prescribed by the social choice function. Weak rationalizability has been introduced by Battigalli (1999, 2003).

by a zero-probability event, Bayesian updating does not apply and the agent needs to revise her belief in another fashion. Precisely the assumption on how this belief revision proceeds is what distinguishes different existing notions of robustness for dynamic mechanisms. Penta (2009) assumes that after a surprise there is common belief that everybody will behave rationally from now on (even if the current node can only be reached by irrational strategies). This corresponds to rationality and common future belief in rationality (Penta, 2009; Perea, 2011). And in Müller (2012) we assume that the agents employ a forward induction logic when revising their beliefs. More specifically, we assume that there is rationality and common strong belief in rationality (Battigalli and Siniscalchi, 2002). In the current paper, by studying weakly rationalizable implementation, we examine which social choice functions are implementable if *no* assumptions whatsoever are made on the belief revision process. This corresponds to assuming rationality and common initial belief in rationality.[2] Rationality and common initial, common future and common strong belief in rationality all collapse to RCBR when applied to static mechanisms. But due to the lack of any belief revision assumption, common initial belief in rationality translates to the weakest of the corresponding solution concepts.[3] And hence, weakly rationalizable implementation is the strongest among the discussed concepts of robust implementation in dynamic mechanisms.

In fact, one might suspect that the absence of any belief revision assumption makes dynamic mechanisms lose their bite. One might question whether under weak rationalizability, dynamic mechanisms can implement more social choice functions than are already rationalizably implementable by static mechanisms. In Müller (2012) we show that dynamic mechanisms can virtually, that is, approximately, implement considerably more social choice functions than static mechanisms (in a robust fashion). But key to this positive result is that by exploiting the forward induction logic embedded in common strong belief in rationality, one can construct a mechanism in which agents "learn" their opponents' payoff types. While weak rationalizability implies that initially, there is common belief in the agents' rationality, as soon as an agent is surprised her beliefs become unrestricted, and she no longer needs to belief in the others' rationality. This prevents "learning" such as that in Müller (2012), as an agent can stubbornly believe in a particular payoff type of her opponent — even if past events contradict that belief. And indeed, in Appendix B we verify that dynamic mechanisms virtually weakly rationalizably implement precisely the same social choice functions as static mechanisms. Maybe surprisingly, our main results imply that when it comes to exact implementation, this is not true. They imply that despite the lack of any belief revision assumption, dynamic mechanisms can (exactly) weakly rationalizably implement *strictly more* social choice functions than static

[2]Compare, e.g., Ben-Porath (1997). See footnote 10 for more details on infinite mechanisms.

[3]Rationality and common strong belief in rationality is characterized by strong rationalizability, and rationality and common future belief in rationality by backwards rationalizability. Every backwards rationalizable and every strongly rationalizable strategy is weakly rationalizable.

mechanisms. Even in situations in which a mechanism designer is not comfortable making any assumption about the agents' belief revision process, dynamic mechanisms can help.

Weak Rationalizability and Weak Perfect Bayesian Equilibrium. As mentioned above, one motivation for studying implementation in weakly rationalizable strategies is based on the relation between weak perfect Bayesian equilibrium and weak rationalizability. While a formal analysis is beyond the scope of this paper, we briefly describe this motivation here. In a typical formulation of a Bayesian game, each player has a payoff type $\theta_i \in \Theta_i$ that is known to her but not to the other players. The state of the world $\theta = (\theta_1, \ldots, \theta_I)$ captures all information that affects the agents' preferences. Two ways have been used to describe an agent's beliefs in such a model. The direct but often not very practical way is to endow agent i with some belief hierarchies, called her epistemic types (for simplicity, let us restrict attention to static games for now). A belief hierarchy for i consists of a first-order belief comprising i's probability measure on Θ_{-i}, a second-order belief comprising i's probability measure on the set of $-i$'s first-order beliefs, an so on, ad infinitum. The set of all belief hierarchies is called the universal type space. The second and indirect way, developed by Harsanyi (1967-68), is familiar from applied game theory. Here, we endow each payoff type θ_i with a probability measure $p_i(\theta_i)$ on Θ_{-i}. The function $\theta_i \mapsto p_i(\theta_i)$ is commonly known among the agents. Often, it is even assumed that $p_i(\theta_i)$ is derived from a common prior p^c on the space of all payoff type profiles $(\theta_1, \ldots, \theta_I)$, so that $p_i(\theta_i) = p^c(\cdot|\theta_i)$. Of course, $p_i(\theta_i)$ corresponds to what was i's first-order belief in the direct approach. But since with θ_{-i} there is associated $-i$'s first-order belief $p_{-i}(\theta_{-i})$, the measure $p_i(\theta_i)$ implicitly also defines a second-order belief for i. Continuing in this fashion, we find that $p_i(\theta_i)$ implicitly defines a whole belief hierarchy. Consequently, exactly one epistemic type corresponds to the payoff type θ_i, namely the belief hierarchy derived from $p_i(\theta_i)$.[4] Implicitly, by following the indirect approach, we allow only specific epistemic types and restrict attention to a "small" type space that is a strict subset of the universal type space.

The concern of the robust implementation literature is that if the mechanism designer makes a mistake in correctly modeling the agents' epistemic types, then the recommended Bayesian mechanism might "malfunction" and not implement the desired social choice function. And this concern is real (see, e.g., Neeman, 2004). Even minuscule changes in the agents' belief hierarchies can lead to different Bayesian equilibrium outcomes (see, e.g., Rubinstein, 1989; Weinstein and Yildiz, 2007). Robustness is introduced by expanding the set of epistemic types that can be associated with a payoff type. In particular, "global" robust implementation demands that a payoff type can be associated with *any* hierarchy of beliefs. Bergemann and

[4]Some formulations of a Bayesian game are more general and allow different epistemic types to be associated with one payoff type (for a more thorough discussion, see Battigalli (1999)). Still, all "small" type space models are non-robust.

Morris (2011), for example, define robust implementation as implementation on every type space (including the universal type space). That is, every outcome that is an equilibrium outcome for some type space has to coincide with the outcome prescribed by the social choice function. A result from epistemic game theory links robust implementation (as just defined) back to rationalizable implementation. The result says that the set of rationalizable outcomes and the union of Bayesian equilibrium outcomes over all type spaces coincide (Brandenburger and Dekel, 1987; Battigalli and Siniscalchi, 2003). Hence, robust implementation is equivalent to rationalizable implementation,[5] and understanding the latter allows us to judge which social choice functions are robustly implementable.

An analogous motivation can be provided for weakly rationalizable implementation by dynamic mechanisms. In dynamic mechanisms, probability measures are replaced by belief systems, describing an agent's belief at each information set. Battigalli and Siniscalchi (1999) formulate a universal type space of such "interactive" beliefs, capturing all hierarchies of belief systems in a dynamic game. And Battigalli (1999) shows the equivalence of weak rationalizability and weakly perfect Bayesian equilibria on all type spaces (in "simple," yet possibly infinite games). Consequently, weakly rationalizable implementation can be motivated as robust implementation in weak perfect Bayesian equilibrium.

Results. Throughout, we work in an environment in which outcomes are lotteries over a finite set of pure outcomes and each agent has finitely many payoff types. Given Bergemann and Morris' (2009b) characterization of robust virtual implementation, showing that (finite) static and (finite) dynamic mechanisms virtually implement exactly the same social choice functions in weakly rationalizable strategies is a comparatively easy task. The set of robustly virtually implementable social choice functions equals — essentially by definition — the set of social choice functions that are virtually weakly rationalizably implementable (virtually wr-implementable) by static mechanisms and is characterized as the set of ex-post incentive compatible (epIC)[6] and robustly measurable social choice functions. Here, Bergemann and Morris (2009b) call a social choice function robustly measurable if it treats strategically indistinguishable payoff types the same. In Appendix B we prove that (under weak rationalizability), dynamic mechanisms can strategically distinguish exactly the same payoff types as static mechanisms (Proposition 4, Corollary 1). This implies that epIC and robust measurability are necessary for virtual wr-implementation (Proposition 5) and we can conclude that

[5] For implementation in infinite mechanisms, the equivalence is not obvious. Implementation requires that every rationalizable outcome (equilibrium outcome) coincides with the outcome specified by the social choice function, and that the set of rationalizable outcomes (equilibrium outcomes) is nonempty. Bergemann and Morris (2011) complete the proof of equivalence of robust and rationalizable implementation by showing that some mechanism satisfies the nonemptiness condition for all type spaces.

[6] On the strength of this incentive compatibility condition see e.g. Jehiel, Meyer-Ter-Vehn, Moldovanu, and Zame (2006), but also Bikhchandani (2006) and Dasgupta and Maskin (2000).

dynamic mechanisms indeed cannot virtually wr-implement more social choice functions than static mechanisms (Corollary 2).

Our main results imply that while dynamic mechanisms virtually wr-implement precisely the same social choice functions as static mechanisms, they exactly weakly rationalizably implement (wr-implement) strictly more social choice functions. Proposition 1 shows that dynamic robust monotonicity (dr-monotonicity) is necessary for wr-implementation. Dr-monotonicity is related to but weaker than Bergemann and Morris' (2011) robust monotonicity condition, central to their characterization of rationalizable implementation. The incentive compatibility condition for wr-implementation is medium-strict ex-post incentive compatibility. This is a version of epIC that is weaker then semi-strict epIC — the incentive compatibility condition that is necessary for rationalizable implementation — but stronger than epIC. Medium-strict epIC does no appear directly in our characterization of wr-implementation as it is implied by dr-monotonicity (Proposition 2).

Proposition 3 shows that under a conditional no total indifference condition, any semi-strict epIC and dr-monotone social choice function is wr-implementable. In order to obtain this clean result, we employ a badly behaved infinite mechanism. However, Examples 3.1 and 3.2 present some social choice functions that are not robustly implementable by a static mechanism but wr-implementable by finite and thus well-behaved dynamic mechanisms. While both the characterization of wr-implementation in well-behaved static and in well-behaved dynamic mechanisms are open questions, we can conclude that well-behaved dynamic mechanisms outperform their static counterparts — in fact, they sometimes even outperform all badly behaved static mechanisms, as well.

Organization of the Paper. Section 2 describes the environment and defines weakly rationalizable implementation. Section 3 derives necessary and Section 4 sufficient conditions for wr-implementation. The results on virtual wr-implementation are relegated to Appendix B. Some readers may want to directly skip ahead to Example 3.1 in Subsection 3.2 and refer back to Section 2 and Subsection 3.1 as necessary. Example 3.1 describes a social choice function which is not robustly implementable by any (finite or infinite) static mechanism, but wr-implementable by a finite dynamic mechanism.

2 Environment and Preliminaries[7]

There is a finite set $\mathcal{I} = \{1, \dots, I\}$ of at least two agents. Each agent $i \in \mathcal{I}$ has a nonempty and finite payoff type space Θ_i. We let Θ denote the set of payoff type profiles $(\theta_1, \dots, \theta_I)$. More generally, if $(Z_i)_{i\in\mathcal{I}}$ is a family of sets Z_i, we let Z denote the Cartesian product $\prod_{i\in\mathcal{I}} Z_i$.

[7]The basic notation is similar to the one in Müller (2012).

It is also understood that z denotes $(z_1, \ldots, z_I)$ whenever $z_i \in Z_i$ for all $i \in \mathcal{I}$, and that $z(\theta)$ denotes $(z_1(\theta_1), \ldots, z_I(\theta_I))$ whenever $z_i \in Z_i^{\Theta_i}$ for all $i \in \mathcal{I}$. If $Z_i = A_i \times B_i$ for all $i \in \mathcal{I}$, we at times ignore the correct order of tuples and write $((a_1, \ldots, a_I), (b_1, \ldots, b_I)) \in Z$ for $(a_i, b_i)_{i \in \mathcal{I}} \in Z$.

There is a nonempty and finite set X of pure outcomes; the set of outcomes is the set $Y = \{y \in \mathbb{R}^{\#X} : y \geq 0, \sum_{n=1}^{\#X} y_n = 1\}$ of lotteries over X. We let $u_i(x, \theta)$ denote the von Neumann-Morgenstern utility that 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. 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 $marg_{\Theta_j} \psi_i$. For $\theta_i \in \Theta_i$, $\delta(\theta_i) \in \Delta(\Theta_i)$ denotes the degenerate belief in θ_i.

2.1 Mechanisms

A (dynamic) mechanism $\Gamma = \langle H, (\mathscr{H}_i)_{i \in \mathcal{I}}, P, C \rangle$ is an extensive game form of finite length, with perfect recall and no trivial decision nodes. The set of dynamic mechanisms includes the set of static mechanisms or normal game forms as a proper subset. We relegate most definitions to Appendix A, but summarize some important notation here. A mechanism's first component, H, is a set of histories $h = (a_1, \ldots, a_n)$, which are finite sequences of actions. We let $\emptyset$ denote the initial history. At any 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 $\{a : (h, a) \in H\}$. Here, (h, a) denotes the history $(a_1, \ldots, a_n, a)$. The set $\mathscr{H}_i$ 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. Once a terminal history h is reached the lottery $C(h) \in Y$ obtains as the outcome of the mechanism.

A strategy s_i for player i specifies an action for each information set $\mathcal{H} \in \mathscr{H}_i$. The set of i's strategies is S_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 $S_i(\mathcal{H})$ be the set of i's strategies admitting j's information set $\mathcal{H}$, $j \in \mathcal{I}$, and $S_{-i}(\mathcal{H})$ be the set of $-i$'s strategies admitting $\mathcal{H}$. For any $\mathcal{J} \subseteq \mathcal{I}$, we let

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

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

2.2 Beliefs and Sequential Rationality

Player i's beliefs about her opponents' strategies and payoff types are captured by a family of probability measures on $(\Sigma_{-i}, \mathcal{B}_{-i})$[8], with each measure representing i's belief at one of her information sets. Player i also holds a belief at the initial history, even if it does not comprise one of her information sets. Formally, i's beliefs are indexed by the members of $\bar{\mathscr{H}}_i = \mathscr{H}_i \cup \{\{\emptyset\}\}$ and form a conditional probability system.

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

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

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

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

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

Condition b) requires that at information set $\mathcal{H}$, agent i places zero (marginal) probability on any strategy of $-i$ which would have prevented that $\mathcal{H}$ occurs. Condition c) says that i uses Bayesian updating "whenever applicable:" Suppose that $\mathcal{H}' \preceq \mathcal{H}$, and that at $\mathcal{H}'$, agent i believes that A will happen with probability $\mu_i(A|\mathcal{H}')$. The play proceeds and i finds herself at $\mathcal{H}$. If $\mathcal{H}$ was no surprise to her — that is, if $\mu_i(\Sigma_{-i}(\mathcal{H})|\mathcal{H}') > 0$ — she now believes in A with probability

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

If, on the other hand, $\mathcal{H}$ did *surprise* her — if $\mu_i(\Sigma_{-i}(\mathcal{H})|\mathcal{H}') = 0$ —, Bayesian updating "does not apply" and condition c) allows any $\mu_i(A|\mathcal{H}) \in [0,1]$, that is, any new estimate of the likelihood of A. We let $\Delta(\Sigma_{-i})$ denote the set of probability measures on Σ_{-i} and $\Delta^{\bar{\mathscr{H}}_i}(\Sigma_{-i})$ denote the set of conditional probability systems on Σ_{-i}. Given a CPS $\mu_i \in \Delta^{\bar{\mathscr{H}}_i}(\Sigma_{-i})$,

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

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 $\mu_i(\cdot|\mathcal{H})$.

[8]The set of actions (see Definition 11 in Appendix A) is metrizable. For each history h, we endow the set of actions $A(h)$ available at h with the relative topology. We endow every finite set (such as Θ_i) with the discrete topology and all product sets (such as S_i and Σ_i) with the product topology. The set $\mathcal{B}_{-i}$ is the Borel σ-algebra on Σ_{-i}.

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 $\mu_i \in \Delta^{\bar{\mathscr{H}}_i}(\Sigma_{-i})$ if for all $\mathcal{H} \in \mathscr{H}_i(s_i)$ and all $s_i' \in S_i(\mathcal{H})$*

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

and both sides of this inequality "make sense." [9]

We let $r_i : \Theta_i \times \Delta^{\bar{\mathscr{H}}_i}(\Sigma_{-i}) \twoheadrightarrow S_i$ 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 $\rho_i : \Delta^{\bar{\mathscr{H}}_i}(\Sigma_{-i}) \twoheadrightarrow \Sigma_i$ 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.

2.3 Weak Rationalizability and WR-Implementation

Battigalli (1999, 2003) defines weak rationalizability for infinite games. We reproduce his definition here.[10]

Definition 3 *For $i \in \mathcal{I}$ let $W_i^0 = \Sigma_i$ and $\Pi_i^0 = \Delta^{\bar{\mathscr{H}}_i}(\Sigma_{-i})$ and recursively define the set W_i^{k+1} of weakly $(k+1)$-rationalizable pairs (s_i, θ_i) for player i by*

$$W_i^{k+1} = \rho_i(\Pi_i^k),$$

and the set Π_i^{k+1} of weakly $(k+1)$-rationalizable beliefs for player i by[11]

$$\Pi_i^{k+1} = \left\{ \mu_i \in \Delta^{\bar{\mathscr{H}}_i}(\Sigma_{-i}) : \mu_i(W_{-i}^{k+1} | \{\emptyset\}) = 1 \right\},$$

$k \in \mathbb{N}$. Finally, let $W_i^\infty = \bigcap_{k=0}^\infty W_i^k$ be the set of weakly rationalizable strategy-payoff type pairs for player i, and $\Pi_i^\infty = \bigcap_{k=0}^\infty \Pi_i^k$ be the set of weakly rationalizable beliefs for player i.

[9] 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. But for some non-measurable functions they might not have their usual interpretation. We thus require that they "make sense." Formally, 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 $\mathcal{B}_{-i}$ completed with respect to $\mu_i(\cdot|\mathcal{H})$. In the remainder of the paper we tacitly use the fact that if $\mu_i(\cdot|\mathcal{H})$, $\mathcal{H} \in \bar{\mathscr{H}}_i$, 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 $\theta_i \in \Theta_i$.

[10] Battigalli (1999, 2003) defines weak rationalizability using (an ordinal number of) ω many rounds of elimination of never-best sequential responses. Lipman (1994) shows that in order to capture common initial belief in rationality in general infinite games one sometimes needs a larger (ordinal) number of rounds of elimination and thus transfinite induction. Adding more rounds of elimination does not affect our results, as our proof of the necessary conditions explicitly constructs a fixed point of the elimination procedure and the iterated elimination procedure of the mechanism employed in the proof of the sufficient conditions converges in finitely many rounds.

[11] We adopt the usual convention that $\mu_i(W_{-i}^{k+1}|\{\emptyset\}) = 1$ is not satisfied if W_{-i}^{k+1} is not measurable.

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. Once an agent is surprised by an information set that she did not expect to occur, she need not believe in her opponents' rationality. For convenience, we let $Q_i^k(\theta_i) = \{s_i \in S_i : (s_i, \theta_i) \in W_i^k\}$ denote the set of weakly (k-)rationalizable strategies for $\theta_i \in \Theta_i$, where $k \in \mathbb{N} \cup \{\infty\}$ and $i \in \mathcal{I}$. By definition, in a static mechanism, a strategy is weakly rationalizable if and only if is rationalizable.

A social choice function $f : \Theta \to Y$ assigns a desired outcome to each payoff type profile. The key to implementing 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 $f(\theta)$. That is, we pursue full implementation in weakly rationalizable strategies. Moreover, we restrict attention to mechanisms whose set of weakly rationalizable strategy profiles is nonempty. In fact, we only consider partially ex-post well-behaved mechanisms.

Definition 4 *Mechanism Γ is partially ex-post well-behaved (pepWB) if there exist nonempty sets $\mathcal{Q}_i(\theta_i) \subseteq Q_i^\infty(\theta_i)$ (where $i \in \mathcal{I}$ and $\theta_i \in \Theta_i$) such that for all $i \in \mathcal{I}$, $\theta \in \Theta$ and $s_{-i} \in \mathcal{Q}_{-i}(\theta_{-i})$, there exist $s_i \in \mathcal{Q}_i(\theta_i)$ and $\mu_i \in \Delta^{\mathscr{H}_i}(\Sigma_{-i})$ such that*

- $s_i \in r_i(\theta_i, \mu_i)$,
- $\mu_i((s_{-i}, \theta_{-i})|\{\emptyset\}) = 1$ *and*
- *for all $\mathcal{H} \in \mathscr{H}_i$ there exists an $s_{-i}^{\mathcal{H}} \in S_{-i}(\mathcal{H})$ such that $\mu_i(\{s_{-i}^{\mathcal{H}}\} \times \Theta_{-i}|\mathcal{H}) = 1$.*

Definition 5 *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) Γ is pepWB.*

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) requires the existence of a sequential best response to some weakly rationalizable CPSs that have a degenerate marginal on the opponents' strategy space. The existence of such sequential best responses simplifies condition (3) of the upcoming definition of d-refutability (Definition 6). To get some idea why this is the case, let us think how we can express j's expected utility $U_j^{\mu_j}(s_j, \theta_j, \mathcal{H})$ in terms of primitives of the environment. If $\mu_j(\{s_{-j}^{\mathcal{H}}\} \times \Theta_{-j}|\mathcal{H}) = 1$ for some $s_{-j}^{\mathcal{H}} \in S_{-j}(\mathcal{H})$ then the "object" that j expects to receive when following the strategy s_j is the lottery $y = C(\zeta(s_j, s_{-j}^{\mathcal{H}})) \in Y$. We can express j's expected utility $U_j^{\mu_j}(s_j, \theta_j, \mathcal{H})$ simply as $E_{\psi_j} u_j(y, \theta)$, where $\psi_j \equiv marg_{\Theta_{-j}} \mu_j(\cdot|\mathcal{H})$. For a general μ_j the "object" that j expects from following s_j is a compound lottery $y' \in Y$. Because j might expect correlations between θ_{-j} and s_{-j}, it does not suffice that we know y' and $marg_{\Theta_{-j}} \mu_j(\cdot|\mathcal{H})$. In this case, to express j's expected utility from s_j we need to know

exactly how much probability j places on each (s_{-j}, θ_{-j}). Thus in the general case, we would need to work with functions $y' : \Theta_{-j} \to Y$ instead of simply lotteries $y \in Y$.

The assumption of pepWB matters. For the reason just mentioned in the previous paragraph, if we would weaken pepWB and just require that $Q_i^\infty(\theta_i) \neq \emptyset$ for all i and θ_i, we would obtain somewhat weaker necessary conditions. Our sufficient conditions would still hold. The more interesting open question seems to be what happens if we strengthen pepWB and restrict attention to well-behaved mechanisms (mechanisms in which every payoff type has some sequential best reply against any CPS). Even if we restrict attention to static mechanisms, this question is unresolved. Considering pepWB mechanisms is a compromise that yields a strong characterization result (Propositions 1 and 3) and allows us to compare our necessary and sufficient conditions to the existing literature (whose results are also based on badly behaved mechanisms).

3 Necessary Conditions for WR-Implementation

In this section, we show that dr-monotonicity is necessary for wr-implementation. Subsection 3.1 introduces the building blocks of dr-monotonicity via "bare-bone" definitions, and compares dr-monotonicity to robust monotonicity. Subsection 3.2 illustrates these definitions in an example. Subsection 3.3 proves the necessity of dr-monotonicity, and Subsection 3.4 shows that the incentive compatibility condition of wr-implementation is already implicit in the dr-monotonicity condition.

3.1 Definitions

First, 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 $\theta_i \in \Theta_i$. It is useful to think of $\beta_i(\theta_i)$ as a set of i's strategies in the direct mechanism associated with the social choice function f under consideration. A deception β is *acceptable* if $\theta' \in \beta(\theta)$ implies $f(\theta') = f(\theta)$ for all $\theta, \theta' \in \Theta$, 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.

Like Bergemann and Morris (2011), we call

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

the reward set for agent i (with respect to θ_{-i}). Moreover, we let

$$\Theta_{-i}^{\theta_i' \leftarrow \theta_i} = \{\theta_{-i}' \in \Theta_{-i} : f(\theta_i, \theta_{-i}') \neq f(\theta')\}$$

be the unacceptable "range" of θ_i's announcement θ_i'.

Definition 6 *A deception β is d-refutable if there exist $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta_i' \in \beta_i(\theta_i)$ such that $\Theta_{-i}^{\theta_i' \leftarrow \theta_i} \neq \emptyset$ and such that for all $\theta_{-i}' \in \Theta_{-i}^{\theta_i' \leftarrow \theta_i}$ and all $\psi_i \in \Delta(\Theta_{-i})$ with $\psi_i(\beta_{-i}^{-1}(\theta_{-i}')) = 1$,*

$$\exists x \in Y_i(\theta_{-i}') : E_{\psi_i} u_i(x, \theta) > E_{\psi_i} u_i(f(\theta'), \theta) \tag{2}$$

$$\begin{aligned} \textit{or} \quad & \exists j \neq i, \theta_j \in supp(\psi_i), \psi_j^{BR} \in \Delta(\Theta_{-j}), y \in Y \; \forall \psi_j \in \Delta(\Theta_{-j}) \; \exists x \in Y : \\ & E_{\psi_j} u_j(x, \theta) > E_{\psi_j} u_j(y, \theta) \textit{ and } E_{\psi_j^{BR}} u_j(x, \theta_j', \theta_{-j}) \leq E_{\psi_j^{BR}} u_j(y, \theta_j', \theta_{-j}) \end{aligned} \tag{3}$$

Definition 7 *Social choice function f is dynamically robustly monotone (dr-monotone) if every unacceptable deception is d-refutable.*

In Proposition 1 we establish that dr-monotonicity is necessary for wr-implementation. Bergemann and Morris (2011, Theorem 1, Corollary 1) show that a related condition, robust monotonicity, is necessary for wr-implementation in static mechanisms.[12] A social choice function is robustly monotone if every unacceptable deception is refutable, where refutability is defined as follows.

Definition 8 *A deception β is refutable if there exist $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta_i' \in \beta_i(\theta_i)$ such that $\Theta_{-i}^{\theta_i' \leftarrow \theta_i} \neq \emptyset$ and such that for all $\theta_{-i}' \in \Theta_{-i}$ and all $\psi_i \in \Delta(\Theta_{-i})$ with $\psi_i(\beta_{-i}^{-1}(\theta_{-i}')) = 1$*

$$\exists x \in Y_i(\theta_{-i}') : E_{\psi_i} u_i(x, \theta) > E_{\psi_i} u_i(f(\theta'), \theta) \tag{4}$$

Note that in order to simplify the comparison with d-refutability, Definition 8 adds the requirement that $\Theta_{-i}^{\theta_i' \leftarrow \theta_i} \neq \emptyset$ to Bergemann and Morris' (2011) original definition of refutability. This is a purely cosmetic change, and Definition 8 equivalent to Bergemann and Morris' (2011) definition of refutability. Clearly, any refutable β is d-refutable, but the reverse is not true. Therefore, dr-monotonicity is weaker than robust monotonicity.

Examples 3.1 and 3.2 will show that there are social choice functions f and deceptions β such that f is wr-implementable (by a dynamic mechanism) and β is unacceptable and not refutable, confirming that if we admit dynamic mechanisms then robust monotonicity is no longer a necessary condition for wr-implementation.

3.2 Example

If f is wr-implementable by a static mechanism then every unacceptable deception is refutable. If we admit dynamic mechanisms, this is no longer the case. Then unacceptable deceptions need only be d-refutable. Example 3.1 tries to provide some intuition for why this is the case.

[12] 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.

D-refutability is weaker than refutability in two regards. First, Definitions 6 and 8 reveal that the condition [(2) or (3)] replaces the condition (4) (which is identical to (2)). Second, refutability requires (4) for all $\theta'_{-i} \in \Theta_{-i}$ and ψ_i, while d-refutability requires [(2) or (3)] only for all $\theta'_{-i} \in \Theta_{-i}^{\theta'_i \leftarrow \theta_i}$ and ψ_i. Example 3.1 focuses on the first of these differences. Specifically, the example's social choice function is injective. In this case, $\Theta_{-i}^{\theta'_i \leftarrow \theta_i} = \Theta_{-i}$ for all i, θ_i, θ'_i and the sole difference between d-refutability and refutability is that [(2) or (3)] replaces (4). We will return to the case that $\Theta_{-i}^{\theta'_i \leftarrow \theta_i} \neq \Theta_{-i}$ for some i, θ_i, θ'_i in Subsection 3.4 and Example 3.2.

While in Subsection 3.3 we will consider both finite and infinite mechanisms, the implementing mechanism in Example 3.1 is finite. Since any finite mechanism Γ is pepWB, we can focus on whether $C(\zeta(s)) = f(\theta)$ for all $(s, \theta) \in W^\infty$ for the purpose of determining whether Γ wr-implements f.

Example 3.1 There are two agents $i \in \{1, 2\}$ with two payoff types each, $\Theta_i = \{\theta_i, \theta'_i\}$. The set of pure outcomes is $X = \{w, w', x, y, z\}$. Figure 1 depicts the social choice function f and the associated direct mechanism Γ^d. Note that in this example, we will use "$\hat{\theta}_i$" and "$\hat{\vartheta}_i$" interchangeably to denote the same payoff type. Generally, we try to use $\hat{\theta}_i$ when we think of $\hat{\theta}_i$ as a payoff type, and $\hat{\vartheta}_i$ if we think of $\hat{\theta}_i$ as a strategy.

f	θ_2	θ'_2
θ_1	w	x
θ'_1	y	z

Γ^d	ϑ_2	ϑ'_2
ϑ_1	w	x
ϑ'_1	y	z

Figure 1: f (left) and associated direct mechanism Γ^d (right)

If both players' preferences over $f(\Theta) = \{w, x, y, z\}$ are such that

$$u_i(f(\hat{\theta}_i, \hat{\theta}'_{-i}), \hat{\theta}) > u_i(f(\hat{\theta}'), \hat{\theta}) \quad \text{for all } i \in \{1, 2\}, \hat{\theta}, \hat{\theta}' \in \Theta \text{ s.t. } \hat{\theta}_i \neq \hat{\theta}'_i$$

then Γ^d wr-implements f (in fact, in dominant strategies). One change, highlighted in red in the tables below, makes implementing f more challenging. We assume that

$$u_i(z, \theta) = u_i(f(\theta'), \theta) > u_i(f(\theta_i, \theta'_{-i}), \theta) \quad \text{for all } i \in \{1, 2\}. \tag{5}$$

We let the "additional" outcome w' always be "slightly" worse than w for player 1, and always be exactly as good as w for player 2. Adding some further assumptions (that will only matter

later) in the right-most column of the following tables, in summary, we assume that

$$\begin{array}{llllllllllll} u_1(\cdot,\theta_1,\theta_2) & \text{represents} & w & \succ & w' & \succ & y, & z & \succ & x, & z & \succ & w, \\ u_1(\cdot,\theta_1,\theta_2') & \text{represents} & w & \succ & w' & \succ & y, & x & \succ & z, & & & \\ u_1(\cdot,\theta_1',\cdot) & \text{represents} & y & \succ & w & \succ & w', & z & \succ & x,^{13} & & & \end{array}$$

and

$$\begin{array}{llllllllllll} u_2(\cdot,\theta_1,\theta_2) & \text{represents} & w & \sim & w' & \succ & x, & z & \succ & y, & w & \succ & z \\ u_2(\cdot,\theta_1',\theta_2) & \text{represents} & w & \sim & w' & \succ & x, & y & \succ & z, & x & \succ & y \\ u_2(\cdot,\cdot,\theta_2') & \text{represents} & x & \succ & w & \sim & w', & z & \succ & y. & & & \end{array}$$

It is clear that (5) prevents Γ^d from wr-implementing f. If payoff type θ_1 believes in (ϑ_2', θ_2), that is, if θ_1 believes she faces a payoff type θ_2 that announces ϑ_2', then θ_1 will announce ϑ_1'. Similarly, the belief in (ϑ_1', θ_1) rationalizes ϑ_2' for θ_2'. Hence ϑ_i' never gets eliminated for θ_i and $f(\theta') = z$ is a weakly rationalizable outcome for the payoff type profile θ.

It is useful to rephrase this insight in terms of deceptions. If $\hat{\beta}$ is some deception, let

$$G_i(\hat{\beta}_i) = \{(\hat{\vartheta}_i', \hat{\theta}_i) \in \Theta_i^2 : \hat{\vartheta}_i' \in \hat{\beta}_i(\hat{\theta}_i)\} \subseteq \Sigma_i^d$$

be the ("inverted") graph of $\hat{\beta}_i$, $i \in \mathcal{I}$, and let

$$\Pi_i(\hat{\beta}) = \Big\{\mu_i \in \Delta^{\bar{\mathscr{H}}_i}(\Sigma_{-i}) : \mu_i\big(\prod_{j\neq i} G_i(\hat{\beta}_j)|\{\emptyset\}\big) = 1\Big\}$$

summarize the CPSs of i that express initial belief in $\prod_{j\neq i} G_i(\hat{\beta}_j)$. We say that $\hat{\beta}$ is a fixed point of the elimination procedure that defines weak rationalizability if $G_i(\hat{\beta}_i) = \rho_i(\Pi_i(\hat{\beta}))$ for all $i \in \mathcal{I}$. If $\hat{\beta}$ is a fixed point, then for all $i \in \{1,2\}$, all $(\hat{\vartheta}_i, \hat{\theta}_i) \in G_i(\hat{\beta}_i)$ survive the iterated elimination of never-best sequential responses, and $G_i(\hat{\beta}_i) \subseteq W_i^{d,\infty}$. Consequently, if Γ^d wr-implements f and $\hat{\beta}$ is a fixed point then $\hat{\vartheta}' \in \hat{\beta}(\hat{\theta})$ implies $C^d(\zeta(\hat{\vartheta}')) = f(\hat{\theta})$. Because $C^d(\zeta(\hat{\vartheta}'))$ is simply $f(\hat{\vartheta}')$ we can conclude that $\hat{\beta}$ must be acceptable. One fixed point of Γ^d's elimination procedure is the deception β such that $\beta_i(\theta_i) = \{\theta_i, \theta_i'\}$ and $\beta_i(\theta_i') = \{\theta_i'\}$ for all $i \in \{1,2\}$. In fact, β is Γ^d's largest fixed point as $G_i(\beta_i) = W_i^{d,\infty}$ for all $i \in \{1,2\}$. We illustrate β in figure 2, where an arrow from $\hat{\theta}_i$ to $\hat{\theta}_i'$ indicates that $\hat{\theta}_i' \in \beta_i(\hat{\theta}_i)$. Unfortunately, β is unacceptable — for example, as highlighted in red in figure 2, $\theta' \in \beta(\theta)$ but $f(\theta') = z \neq w = f(\theta)$ — and thus implies that Γ^d does not wr-implement f.

Next, let us consider the possibility of wr-implementing f by an indirect static mechanism Γ^s. If Γ^s wr-implements f, then for each $\hat{\theta} \in \Theta$ there exists a strategy profile $\hat{s} \in Q^{s,\infty}(\hat{\theta})$.

[13] That is, $u_1(\cdot,\theta) : X \to \mathbb{R}$ represents some preference relation $\succeq$ such that $w \succ w' \succ y$, $z \succ x$ and $z \succ w$, where $\succ$ is the strict preference relation derived from $\succeq$. The last line of the table says that both $u_1(\cdot,\theta_1',\theta_2) : X \to \mathbb{R}$ and $u_1(\cdot,\theta_1',\theta_2') : X \to \mathbb{R}$ represent some preference relation satisfying $y \succ w \succ w'$ and $z \succ x$.

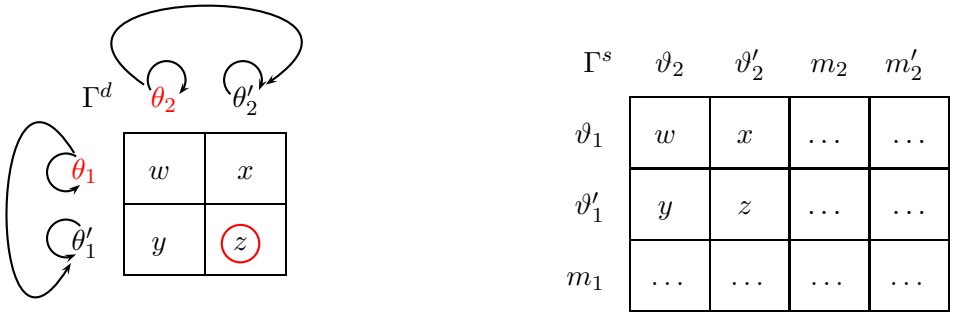

Figure 2: Direct mechanism Γ^d (left) and augmented direct mechanism Γ^s (right)

Without loss of generality, we can rename $\hat{s}$ and henceforth denote it by $\hat{\vartheta} \in \Theta$. After this relabeling, we can write i's strategy set as $S_i^s = \Theta_i \cup M_i$, where M_i is a (potentially empty) set of messages $m_i \notin \Theta_i$. As illustrated in figure 2, we have $C^s(\zeta(\hat{\vartheta})) = f(\hat{\theta})$ for every $\hat{\vartheta} \in \Theta$, which justifies calling Γ^s an "augmented direct mechanism."

As above, if Γ^s wr-implements f, then any unacceptable deception $\hat{\beta}$ — in particular β — must not form a fixed point of Γ^s's iterated elimination procedure. This implies that for some $i \in \{1,2\}$ and $\hat{\theta}_i \in \Theta_i$, some $\hat{\vartheta'}_i \in \beta_i(\hat{\theta}_i)$ must be eliminated if $\hat{\theta}_i$'s beliefs are restricted to $\Pi_i(\beta)$. That is, we must have $\hat{\vartheta'}_i \notin r_i(\hat{\theta}_i, \mu_i)$ for i, $\hat{\theta}_i$, $\hat{\vartheta'}_i$ and any $\mu_i \in \Pi_i(\beta)$, and in particular for any μ_i such that

$$\mu_i((\hat{\vartheta}_j', \hat{\theta}_j)|\{\emptyset\}) = \psi_i(\hat{\theta}_j) \quad \text{for all } \hat{\theta}_j \in \Theta_j$$

for some $\hat{\vartheta}_j' \in \Theta_j$ and some $\psi_i \in \Delta(\Theta_j)$ with $\psi_i(\beta_j^{-1}(\hat{\vartheta}_j')) = 1$. Hence for any $\hat{\vartheta}_j' \in \Theta_j$ and any $\psi_i \in \Delta(\Theta_j)$ with $\psi_i(\beta_j^{-1}(\hat{\vartheta}_j')) = 1$ there must be some strategy m_i such that

$$E_{\psi_i} u_i(C^s(\zeta(m_i, \hat{\vartheta}_j')), \hat{\theta}) > E_{\psi_i} u_i(C^s(\zeta(\hat{\vartheta}')), \hat{\theta}).$$

Recalling that $C^s(\zeta(\hat{\vartheta}')) = f(\hat{\vartheta}')$, we can conclude that for any $\hat{\vartheta}_j'$ and ψ_i there must exist an $a \in Y$ — namely $a = C^s(\zeta(m_i, \hat{\vartheta}_j'))$ — such that

$$E_{\psi_i} u_i(a, \hat{\theta}) > E_{\psi_i} u_i(f(\hat{\vartheta}'), \hat{\theta}). \tag{6}$$

Let us think about which $\hat{\vartheta'}_i \in \beta_i(\hat{\theta}_i)$ can possibly be eliminated. Start by considering $\vartheta_i' \in \beta_i(\theta_i)$ for $i \in \{1,2\}$. For $i = 1$ we have

$$u_i(a, \theta) \leq u_i(z, \theta) = u_i(f(\vartheta'), \theta) \quad \text{for all } a \in Y.$$

Hence for ϑ_2' and $\psi_1 = \delta(\theta_2)$ there is no $a \in Y$ that satisfies (6), and $\vartheta_1' \in \beta_1(\theta_1)$ cannot be eliminated. For $i = 2$, on the other hand, there appears to be more hope. Let us focus on ϑ_1' and $\psi_2 = \delta(\theta_1)$. Then (6) simplifies to $u_2(a, \theta) > u_2(f(\vartheta'), \theta) = u_2(z, \theta)$ and is satisfied

by $a = w'$.[14] Hence we can dissuade a payoff type θ_2 with belief (ϑ_1', θ_1) from playing ϑ_2' by augmenting Γ^d as shown on the left in figure 3 (θ_2's belief is illustrated by the dashed arrow). Such a θ_2 will now play m_2 instead of ϑ_2'.

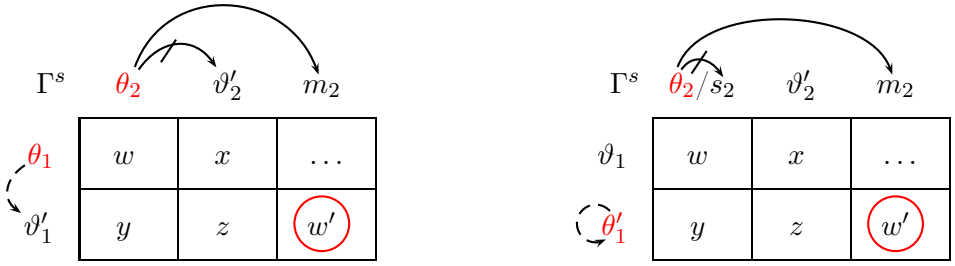

Figure 3: Augmented direct mechanism Γ^s

However, this augmentation introduces another problem. If Γ^s wr-implements f, then θ_2 has some best response s_2 against the belief that she faces a truth-telling payoff type θ_1'. But given this belief, θ_2 strictly prefers w' over $C^s(\zeta(\vartheta_1', s_2)) = f(\vartheta_1', \vartheta_2) = y$, and hence m_2 over s_2 (righthand side of figure 3). Contradiction. Hence, an additional requirement on the a satisfying (6) is that

$$u_2(a, \theta_1', \theta_2) \leq u_2(f(\vartheta_1', \theta_2), \theta_1', \theta_2) = u_2(y, \theta_1', \theta_2).$$

In fact, repeating the last argument for θ_2', we see that we need $a \in Y_2(\theta_1')$. Since $Y_2(\theta_1')$ only contains y, z and all mixtures of y and z, no $a \in Y$ satisfies both (6) and $a \in Y_2(\theta_1')$. Hence we cannot eliminate $\vartheta_2' \in \beta_2(\theta_2)$ either.

Summing up, we argued that if Γ^s wr-implements f and $\hat{\beta}$ is unacceptable, then there is $i \in \mathcal{I}$, $\hat{\theta}_i \in \Theta_i$ and $\hat{\vartheta}_i' \in \hat{\beta}_i(\hat{\theta}_i)$ such that for all $\hat{\vartheta}_j' \in \Theta_j$ and all $\psi_i \in \Delta(\Theta_j)$ with $\psi_i(\hat{\beta}_j^{-1}(\hat{\vartheta}_j')) = 1$

$$\exists a \in Y_i(\hat{\vartheta}_j') : E_{\psi_i} u_i(a, \hat{\theta}) > E_{\psi_i} u_i(f(\hat{\vartheta}'), \hat{\theta}). \tag{7}$$

But this just says that any unacceptable $\hat{\beta}$ must be refutable (recall that $\Theta_j^{\hat{\vartheta}_i' \leftarrow \hat{\theta}_i} = \Theta_j$ as f is injective) — and f robustly monotone. For β, we showed that $\vartheta_i' \in \beta_i(\theta_i)$ cannot be refuted for any $i \in \{1, 2\}$. Since telling the truth can neither be refuted for any payoff type[15], β is not refutable and hence a fixed point of Γ^s's elimination procedure. Consequently, no static Γ^s wr-implements f.

Finally, let us consider a dynamic mechanism Γ. Suppose that Γ wr-implements f and that $\hat{\beta}$ is unacceptable. As above, we can relabel some $\hat{s} \in Q^\infty(\hat{\theta})$ as $\hat{\vartheta}$ for each $\hat{\theta} \in \Theta$, interpret $\hat{\beta}$ as a correspondence mapping payoff type profiles to strategy profiles and conclude that $\hat{\beta}$

[14] Condition (6) is also satisfied by $a = w$ and potentially by $a = x$, and by any mixture of w and w' and (potentially) x. We will see in a moment that this does not matter.

[15] Let $i \in \{1, 2\}$, $\hat{\theta}_i \in \Theta_i$, $\hat{\vartheta}_i = \hat{\theta}_i \in \beta_i(\hat{\theta}_i)$, $\hat{\vartheta}_j \in \Theta_j$ and $\psi_i = \delta(\hat{\vartheta}_j)$. Then if a satisfies (6), $a \notin Y_i(\hat{\vartheta}_j)$.

cannot be a fixed point of Γ's elimination procedure. This time, we will want to extend this conclusion to deceptions that are "equivalent" to $\hat{\beta}$. Since Γ is dynamic, it is conceivable that a strategy $\hat{\vartheta}_i'$ gets eliminated for $\hat{\theta}_i$ simply because it is suboptimal at some information set $\mathcal{H}$ that is not admitted by any weakly rationalizable strategy profile. If $\hat{\beta}_i(\hat{\theta}_i)$ contains such a $\hat{\vartheta}_i'$ then $\hat{\beta}$ is not a fixed point. But it might be that redefining $\hat{\vartheta}_i'$ at $\mathcal{H}$ (but nowhere else) makes $\hat{\vartheta}_i'$ a sequential best response for $\hat{\theta}_i$. Clearly, the "equivalent" version of $\hat{\beta}$ with the redefined instead of the original $\hat{\vartheta}_i'$ must not be a fixed point either.

This is why we consider the following "translations" $\hat{\beta}^t$ of $\hat{\beta}$. Let $\tilde{\vartheta}, \tilde{\theta} \in \Theta$ be such that $\tilde{\vartheta} \in \hat{\beta}(\tilde{\theta})$ and $f(\tilde{\vartheta}) \neq f(\tilde{\theta})$. Let $s^{\tilde{\vartheta}} \equiv (s_i^{\tilde{\vartheta}_i})_{i \in \mathcal{I}} \in Q^\infty(\tilde{\vartheta})$. For each $i \in \mathcal{I}$ let $\hat{\beta}_i^t : \Theta_i \twoheadrightarrow S_i$ be nonempty-valued and such that for any $\hat{\theta}_i \in \Theta_i$,

$$\tilde{\vartheta}_i \in \hat{\beta}_i(\hat{\theta}_i) \quad \text{implies} \quad \hat{\beta}_i^t(\hat{\theta}_i) \cap \left\{ s_i \in S_i : s_i(\mathcal{H}) = s_i^{\tilde{\vartheta}_i}(\mathcal{H}) \text{ for all } \mathcal{H} \in \mathscr{H}_i(s^{\tilde{\vartheta}}) \right\} \neq \emptyset.$$

No such "translation" $\hat{\beta}^t$ of $\hat{\beta}$ can be a fixed point of Γ's elimination procedure.[16]

Now return to β and recall that $\vartheta' \in \beta(\theta)$ and $f(\vartheta') \neq f(\theta)$. Consider the dynamic mechanism Γ depicted in figure 4. Inspecting the assignment of outcomes to terminal histories, we see that if Γ wr-implements f then any weakly rationalizable strategy of θ_i', $i \in \{1, 2\}$, prescribes the action $\vartheta_i'^a$ at i's initial information set. Let us check if some β^t such that for

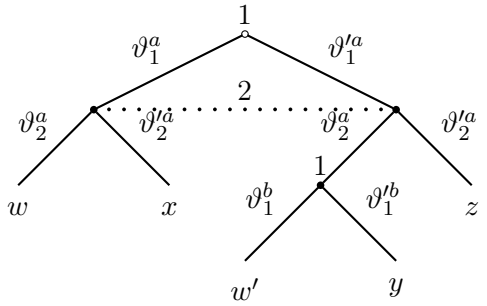

Figure 4: Mechanism Γ that wr-implements f

each $i \in \mathcal{I}$ and $\hat{\theta}_i \in \Theta_i$

$$\beta^t(\hat{\theta}_i) \cap \left\{ s_i \in S_i : s_i(\mathcal{H}) = \vartheta_i'^a \text{ for all } \mathcal{H} \in \mathscr{H}_i \cap \left\{ \{\emptyset\}, \{\vartheta_1^a, \vartheta_1'^a\} \right\} \right\} \neq \emptyset$$

is a fixed point of Γ's elimination procedure. If yes, then Γ does not wr-implement f, as some strategy profile in $\beta^t(\theta)$ survives the elimination procedure and leads to $z = f(\vartheta')$.

Consider $i = 2$. Suppose that θ_2 initially believes that she faces a payoff type θ_1 that "pretends" to be θ_1' on path to $f(\vartheta') = z$. Note that player 1 chooses between w' and y after

[16] For each $\hat{\theta}_i \in \hat{\beta}_i^{-1}(\tilde{\vartheta}_i)$ let $s_i^{\tilde{\vartheta}_i \leftarrow \hat{\theta}_i} \in \hat{\beta}_i^t(\hat{\theta}_i)$ be such that $s_i^{\tilde{\vartheta}_i \leftarrow \hat{\theta}_i}(\mathcal{H}) = s_i^{\tilde{\vartheta}_i}(\mathcal{H})$ for all $\mathcal{H} \in \mathscr{H}_i(s^{\tilde{\vartheta}})$. If $\hat{\beta}^t$ is a fixed point, then $f(\tilde{\vartheta}) = C(\zeta(s^{\tilde{\vartheta}})) = C(\zeta(s^{\tilde{\vartheta} \leftarrow \tilde{\theta}})) = f(\tilde{\theta})$. Contradiction.

the history $(\vartheta_1'^a, \vartheta_2^a)$. Since θ_1 always prefers w' over y and θ_1' always y over w', if β^t is a fixed point then all strategies in $\beta_1^t(\theta_1)$ prescribe ϑ_1^b and all strategies in $\beta_1^t(\theta_1')$ prescribe $\vartheta_1'^b$. Hence, if θ_2 believes to face a payoff type θ_1 that pretends to be θ_1' on path to z (a θ_1 playing $\vartheta_1'^a$), she must belief that "off path," θ_1 will play ϑ_1^b. This gives θ_2 the incentive to deviate from the lie $\vartheta_2'^a$ and play ϑ_2^a instead: θ_2 expects that playing ϑ_2^a leads to w', which a θ_2 believing in θ_1 strictly prefers to z. Since θ_2 also strictly prefers ϑ_2^a over $\vartheta_2'^a$ if she believes in a truth-telling player 1, $\vartheta_2'^a$ gets eliminated from $\beta^t(\theta_2)$. No β^t is a fixed point.

There are other dynamic mechanisms for which no β^t is a fixed point either (for example, the infinite mechanism of Proposition 3). What makes Γ similar to all those mechanisms is that at the information set $\{(\vartheta_1'^a, \vartheta_2^a)\}$, payoff type θ_1' chooses $\vartheta_1'^b$ for some belief $\psi_1^{BR} \in \Delta(\Theta_2)$ — in fact, in this example, for all such beliefs. The resulting outcome, y, is strictly worse for θ_1 than some other outcome that player 1 can bring about at $\{(\vartheta_1'^a, \vartheta_2^a)\}$, namely w'. This is true independent of θ_1's belief $\psi_1 \in \Delta(\Theta_2)$ at $\{(\vartheta_1'^a, \vartheta_2^a)\}$. More formally, observe that for $i = 2$, θ_2 and $\theta_2' \in \beta_2(\theta_2)$, for θ_1' and $\psi_2 = \delta(\theta_1)$,

$$\exists \psi_1^{BR} \in \Delta(\Theta_2),\, a(=y) \in Y\ \forall \psi_1 \in \Delta(\Theta_2)\ \exists b(=w') \in Y:$$
$$E_{\psi_1} u_1(b,\theta) > E_{\psi_1} u_1(a,\theta) \text{ and } E_{\psi_1^{BR}} u_1(b,\theta_1',\theta_2) \leq E_{\psi_1^{BR}} u_1(a,\theta_1',\theta_2).$$

This is of course an instance of (3), and the reason why β is d-refutable but not refutable.

To conclude the example, note that Γ does indeed wr-implement f. The elimination of never-best sequential responses proceeds as follows.

- $Q_1^1(\theta_1) = \{\vartheta_1^a \cdot, \vartheta_1'^a \vartheta_1^b\}$, $Q_1^1(\theta_1') = \{\vartheta_1'^a \vartheta_1'^b\}$,
  $Q_2^1(\theta_2) = \{\vartheta_2^a, \vartheta_2'^a\}$, $Q_2^1(\theta_2') = \{\vartheta_2'^a\}$,
- $Q_1^2(\theta_1) = \{\vartheta_1^a \cdot, \vartheta_1'^a \vartheta_1^b\}$, $Q_1^2(\theta_1') = \{\vartheta_1'^a \vartheta_1'^b\}$,
  $Q_2^2(\theta_2) = \{\vartheta_2^a\}$, $Q_2^2(\theta_2') = \{\vartheta_2'^a\}$,
- $Q_1^3(\theta_1) = \{\vartheta_1^a \cdot\}$, $Q_1^3(\theta_1') = \{\vartheta_1'^a \vartheta_1'^b\}$,
  $Q_2^3(\theta_2) = \{\vartheta_2^a\}$, $Q_2^3(\theta_2') = \{\vartheta_2'^a\}$.

3.3 Necessary Conditions

We now prove that dr-monotonicity is necessary for wr-implementation. Supposing that there is an unacceptable but not d-refutable deception, the proof constructs a fixed point of the iterated elimination procedure defining weak rationalizability. The proof exploits that any implementing mechanism is pepWB.

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

Proof. Suppose that mechanism Γ wr-implements f and that β is an unacceptable deception. Suppose by contradiction that β is not d-refutable. For any $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\theta_i' \in \beta_i(\theta_i)$ such that $\Theta_{-i}^{\theta_i' \leftarrow \theta_i} \neq \emptyset$ let $\theta_{-i}^{\theta_i' \leftarrow \theta_i} \in \Theta_{-i}^{\theta_i' \leftarrow \theta_i}$ 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

$$\forall x \in Y_i(\theta_{-i}^{\theta_i' \leftarrow \theta_i}) : \quad E_{\psi_i^{\theta_i' \leftarrow \theta_i}} u_i(x, \theta) \leq E_{\psi_i^{\theta_i' \leftarrow \theta_i}} u_i(f(\theta_i', \theta_{-i}^{\theta_i' \leftarrow \theta_i}), \theta) \tag{8}$$

and

$$\begin{aligned} \forall j \neq i,\, \theta_j \in supp(\psi_i^{\theta_i' \leftarrow \theta_i}), \psi_j \in \Delta(\Theta_{-j}), y \in Y\ \exists \psi_{\theta_j \Diamond y}^{\theta_i' \leftarrow \theta_i} \in \Delta(\Theta_{-j})\ \forall x \in Y : \\ E_{\psi_{\theta_j \Diamond y}^{\theta_i' \leftarrow \theta_i}} u_j(x, \theta) \leq E_{\psi_{\theta_j \Diamond y}^{\theta_i' \leftarrow \theta_i}} u_j(y, \theta) \text{ or } E_{\psi_j} u_j(x, \theta_j', \theta_{-j}) > E_{\psi_j} u_j(y, \theta_j', \theta_{-j}). \end{aligned} \tag{9}$$

For each $\theta \in \Theta$, fix a $s(\theta) \in \mathcal{Q}^\infty(\theta) \subseteq Q^\infty(\theta)$. Without loss of generality, for all $i \in \mathcal{I}$ and $\theta_i \in \Theta_i$, let $s_i(\theta_i)$ be a sequential best response to some $\mu_i(\theta_i) \in \Pi_i^\infty$ such that for each $\mathcal{H} \in \mathscr{H}_i$, $\mu_i(\theta_i)(\{\hat{s}_{-i}^{\theta_i, \mathcal{H}}\} \times \Theta_{-i} | \mathcal{H}) = 1$ for some $\hat{s}_{-i}^{\theta_i, \mathcal{H}} \in S_{-i}(\mathcal{H})$. This is possible because Γ is pepWB. For convenience, let

$$N = \{(i, \theta_i, \theta_i') \in \mathcal{I} \times \Theta_i^2 : \theta_i' \in \beta_i(\theta_i) \text{ and } \Theta_{-i}^{\theta_i' \leftarrow \theta_i} \neq \emptyset\}.$$

Colloquially, we might say that $(i, \theta_i, \theta_i') \in N$ if θ_i' is a lie of θ_i "that matters" and is permitted under β.

Step 1. Suppose that $(i, \bar{\theta}_i, \bar{\theta}_i') \in N$. We claim that there exists a strategy $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \in Q_i^1(\bar{\theta}_i)$ such that $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\mathcal{H}) = s_i(\bar{\theta}_i')(\mathcal{H})$ for all $\mathcal{H} \in \mathscr{H}_i$ containing a history in

$$\begin{aligned} H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} = \big\{ h \in H : h \preceq \zeta(s_i(\bar{\theta}_i'), s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})) \text{ or } h \preceq \zeta(s_j, s_{-j}(\theta_{-j}^{\theta_j' \leftarrow \theta_j})) \text{ for some} \\ (j, \theta_j, \theta_j') \in N, s_j \in S_j \text{ s.t. } j \neq i, \theta_i^{\theta_j' \leftarrow \theta_j} = \bar{\theta}_i' \text{ and } \bar{\theta}_i \in supp(\psi_j^{\theta_j' \leftarrow \theta_j}) \big\}. \end{aligned}$$

Loosely speaking, we claim that $\bar{\theta}_i$ has a sequentially rational strategy $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}$ that coincides with $s_i(\bar{\theta}_i')$ on path to $\zeta(s_i(\bar{\theta}_i'), s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))$, and "wherever any lie $\theta_j' \in \beta_j(\theta_j)$ of j that matters relies on $\bar{\theta}_i$'s lie $\bar{\theta}_i'$." Let μ_i be a CPS that satisfies the following conditions.

1. For all $\theta_{-i} \in \Theta_{-i}$, $\mu_i((s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}), \theta_{-i}) | \{\emptyset\}) = \psi_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\theta_{-i})$.

2. Suppose that $\mathcal{H} \in \mathscr{H}_i$ is a surprise given the beliefs μ_i (that is, suppose that if $\mathcal{H}^p$ denotes $\mathcal{H}$'s immediate predecessor in $\mathscr{H}_i$ then $\mu_i(\Sigma_{-i}(\mathcal{H}) | \mathcal{H}^p) = 0$), and that $\mathcal{H} \cap H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \neq \emptyset$. Then for some $(j, \theta_j, \theta_j') \in N$ and some $s_j \in S_j$ such that $j \neq i$, $\theta_i^{\theta_j' \leftarrow \theta_j} = \bar{\theta}_i'$ and $\bar{\theta}_i \in supp(\psi_j^{\theta_j' \leftarrow \theta_j})$, $\mathcal{H}$ contains a predecessor of $\zeta(s_j, s_{-j}(\theta_{-j}^{\theta_j' \leftarrow \theta_j}))$. Let $y = C(\zeta(s_i(\bar{\theta}_i'), \hat{s}_{-i}^{\bar{\theta}_i', \mathcal{H}}))$. That is, y is the outcome that $\bar{\theta}_i'$ expects at $\mathcal{H}$ under $\mu_i(\bar{\theta}_i')$ if she follows the strategy

$s_i(\bar{\theta}_i')$. Moreover, let $\bar{\psi}_i = marg_{\Theta_{-i}}\mu_i(\bar{\theta}_i')(\cdot|\mathcal{H})$. We require from μ_i that for some $\psi^{\theta_j' \leftarrow \theta_j}_{\bar{\theta}_i \Diamond y}$ satisfying (9) for i, $\bar{\theta}_i$, $\bar{\psi}_i$ and y,

$$\mu_i((\hat{s}_{-i}^{\bar{\theta}_i', \mathcal{H}}, \theta_{-i})|\mathcal{H}) = \psi^{\theta_j' \leftarrow \theta_j}_{\bar{\theta}_i \Diamond y}(\theta_{-i}) \quad \text{for all } \theta_{-i} \in \Theta_{-i}.$$

3. Suppose that $\mathcal{H} \in \mathscr{H}_i$ is a surprise given the beliefs μ_i, and that $\mathcal{H} \cap H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} = \emptyset$. Then we require that $\mu_i(\cdot|\mathcal{H})$ equals $\mu_i'(\cdot|\mathcal{H})$ for some CPS μ_i' whose marginal on S_{-i} is degenerate for all $\mathcal{H}' \in \mathscr{H}_i$ and against which $\bar{\theta}_i$ has a sequential best response (such a CPS exists because Γ is pepWB). In fact, if possible, choose μ_i' such that some sequential best response of $\bar{\theta}_i$ against μ_i' admits $\mathcal{H}$.

Then some $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \in r_i(\bar{\theta}_i, \mu_i)$ satisfies $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\mathcal{H}) = s_i(\bar{\theta}_i')(\mathcal{H})$ for all $\mathcal{H} \in \mathscr{H}_i$ such that $\mathcal{H} \cap H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \neq \emptyset$. To see this, note the following.

4. 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^{BR} for θ_i'' against μ_i'. Such a μ_i' exists because Γ is pepWB. On the one hand, we must have $C(\zeta(s_i^{BR}, s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))) \neq f(\theta_i'', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})$, 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." On the other hand, $C(\zeta(s_i^{BR}, s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))) = f(\theta_i'', \theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i})$ as $s_i^{BR} \in Q_i^\infty(\theta_i'')$ and Γ wr-implements f. 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 (8),

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

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

5. For a surprise $\mathcal{H} \in \mathscr{H}_i(s_i(\bar{\theta}_i'))$ such that $\mathcal{H} \cap H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \neq \emptyset$ take the corresponding $j \neq i$, θ_j, $\theta_j' \in \beta_j(\theta_j)$, s_j and y and $\bar{\psi}_i$ from condition 2. above. Then by (9), for any $x \in Y$

$$E_{\psi^{\theta_j' \leftarrow \theta_j}_{\bar{\theta}_i \Diamond y}} u_i(x, \bar{\theta}_i, \theta_{-i}) \leq E_{\psi^{\theta_j' \leftarrow \theta_j}_{\bar{\theta}_i \Diamond y}} u_i(y, \bar{\theta}_i, \theta_{-i}) = U_i^{\mu_i}(s_i(\bar{\theta}_i'), \bar{\theta}_i, \mathcal{H}) \tag{10}$$

$$\text{or} \quad E_{\bar{\psi}_i} u_i(x, \bar{\theta}_i', \theta_{-i}) > E_{\bar{\psi}_i} u_i(y, \bar{\theta}_i', \theta_{-i}) = U_i^{\mu_i(\bar{\theta}_i')}(s_i(\bar{\theta}_i'), \bar{\theta}_i', \mathcal{H}). \tag{11}$$

Take $x \in Y$ and suppose that $x = C(\zeta(s_i, \hat{s}_{-i}^{\bar{\theta}_i', \mathcal{H}}))$ for some $s_i \in S_i(\mathcal{H})$. That is, suppose that $\bar{\theta}_i'$ believes at $\mathcal{H}$ that she can bring about x (by following s_i). Then $E_{\bar{\psi}_i} u_i(x, \bar{\theta}_i', \theta_{-i}) = U_i^{\mu_i(\bar{\theta}_i')}(s_i, \bar{\theta}_i', \mathcal{H})$ and (11) contradicts $s_i(\bar{\theta}_i')$ being utility maximizing for $\bar{\theta}_i'$ at $\mathcal{H}$. Consequently, (10) must hold for any x such that $x = C(\zeta(s_i, \hat{s}_{-i}^{\bar{\theta}_i', \mathcal{H}}))$ for some $s_i \in S_i(\mathcal{H})$, and $s_i(\bar{\theta}_i')$ maximizes $\bar{\theta}_i$'s expected utility with respect to $\mu_i(\cdot|\mathcal{H})$ at $\mathcal{H}$.

We can construct the desired $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \in r_i(\bar{\theta}_i, \mu_i)$ as follows.

- Let $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\mathcal{H}') = s_i(\bar{\theta}_i')(\mathcal{H}')$ for all $\mathcal{H}' \in \mathscr{H}_i(s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))$.

  [Note that independently of how we will complete the definition of $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}$ below, we can already say that then by 4., $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}$ maximizes $\bar{\theta}_i$'s expected utility with respect to $\mu_i(\cdot|\mathcal{H}')$ for all $\mathcal{H}' \in \mathscr{H}_i(s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}, s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))$.]

- If $\mathcal{H} \in \mathscr{H}_i$, $\mathcal{H} \cap H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \neq \emptyset$, $\mathcal{H} \notin \mathscr{H}_i(s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))$ and $\mathcal{H}^p \in \mathscr{H}_i(s_{-i}(\theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}))$ where $\mathcal{H}^p$ denotes $\mathcal{H}$'s immediate predecessor in $\mathscr{H}_i$, then let $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\mathcal{H}') = s_i(\bar{\theta}_i')(\mathcal{H}')$ for all $\mathcal{H}' \in \mathscr{H}_i(\hat{s}_{-i}^{\bar{\theta}_i', \mathcal{H}})$ such that $\mathcal{H} \preceq \mathcal{H}'$.

  [Take such an $\mathcal{H}'$ and suppose that $\mathcal{H}' \in \mathscr{H}_i(s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}, \hat{s}_{-i}^{\bar{\theta}_i', \mathcal{H}})$. That is, suppose that $\mathcal{H}'$ is not prevented by $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}$ itself. Then $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}$ maximizes $U_i^{\mu_i}(\cdot, \bar{\theta}_i, \mathcal{H}')$ by 5., independently of how we will complete the definition of $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}$ below.]

- Repeat the last step for all $\mathcal{H} \in \mathscr{H}_i$ such that $\mathcal{H} \cap H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \neq \emptyset$ for which we have not defined $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\mathcal{H})$ yet, but for which we have already defined $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\mathcal{H}^p)$. Iterate this procedure until no such such information sets $\mathcal{H}$ remain.

- All remaining information sets $\mathcal{H}'$ must be such that $\mathcal{H}' \cap H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} = \emptyset$. For these $\mathcal{H}'$, let $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\mathcal{H}')$ equal (one of) the sequential best response(s) mentioned in 3.

Step 2. Suppose that $i \in \mathcal{I}$, $\bar{\theta}_i$, $\bar{\theta}_i' \in \beta_i(\bar{\theta}_i)$ and $\Theta_{-i}^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} = \emptyset$. We claim that there exists a strategy $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} \in Q_i^\infty(\bar{\theta}_i)$ such that $s_i^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}(\mathcal{H}) = s_i(\bar{\theta}_i')(\mathcal{H})$ for all $\mathcal{H} \in \mathscr{H}_i$ containing a history in

$$H_o^{\bar{\theta}_i' \leftarrow \bar{\theta}_i} = \big\{h \in H : h \preceq \zeta(s_j, s_{-j}(\theta_{-j}^{\theta_j' \leftarrow \theta_j})) \text{ for some } (j, \theta_j, \theta_j') \in N, s_j \in S_j \text{ s.t. } j \neq i, \theta_i^{\theta_j' \leftarrow \theta_j} = \bar{\theta}_i' \text{ and } \bar{\theta}_i \in supp(\psi_j^{\theta_j' \leftarrow \theta_j})\big\}. \tag{12}$$

Let μ_i be a CPS that satisfies the following conditions.

- First, $\mu_i(\cdot|\{\emptyset\}) = \mu_i(\bar{\theta}_i)(\cdot|\{\emptyset\})$.

- Conditions 2. and 3. from above, with $H^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}$ replaced by $H_o^{\bar{\theta}_i' \leftarrow \bar{\theta}_i}$ in both cases.

Then $\mu_i \in \Pi_i^\infty$. Some sequential best response $s_i^{\bar\theta_i' \leftarrow \bar\theta_i}$ of $\bar\theta_i$ against μ_i satisfies $s_i^{\bar\theta_i' \leftarrow \bar\theta_i}(\mathcal{H}) = s_i(\bar\theta_i')(\mathcal{H})$ for all $\mathcal{H} \in \mathscr{H}_i$ such that $\mathcal{H} \cap H_o^{\bar\theta_i' \leftarrow \bar\theta_i} \neq \emptyset$. To see this, first consider the information set $\mathcal{H} \in \mathscr{H}_i(\hat{s}_{-i}^{\bar\theta_i,\{\emptyset\}})$ that has no predecessor in $\mathscr{H}_i$ — that is, consider i's first information set in $\mathscr{H}_i(\hat{s}_{-i}^{\bar\theta_i,\{\emptyset\}})$ (if no such $\mathcal{H}$ exists because $\mathscr{H}_i(\hat{s}_{-i}^{\bar\theta_i,\{\emptyset\}}) = \emptyset$ then skip this part of the proof). At $\mathcal{H}$, i holds the belief $\mu_i(\cdot|\{\emptyset\})$. Since $\Theta_{-i}^{\bar\theta_i' \leftarrow \bar\theta_i} = \emptyset$, we have $f(\bar\theta_i, \theta_{-i}) = f(\bar\theta_i', \theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$ and

$$\begin{aligned} U_i^{\mu_i(\bar\theta_i)}(s_i(\bar\theta_i), \bar\theta_i, \{\emptyset\}) &= E_{marg_{\Theta_{-i}}\mu_i(\bar\theta_i)(\cdot|\{\emptyset\})} u_i(f(\bar\theta_i, \theta_{-i}), \bar\theta_i, \theta_{-i}) \\ &= E_{marg_{\Theta_{-i}}\mu_i(\bar\theta_i)(\cdot|\{\emptyset\})} u_i(f(\bar\theta_i', \theta_{-i}), \bar\theta_i, \theta_{-i}) = U_i^{\mu_i(\bar\theta_i)}(s_i(\bar\theta_i'), \bar\theta_i, \{\emptyset\}). \end{aligned}$$

By definition of $s_i(\bar\theta_i)$ and $\mu_i(\bar\theta_i)$,

$$U_i^{\mu_i(\bar\theta_i)}(s_i(\bar\theta_i'), \bar\theta_i, \{\emptyset\}) = U_i^{\mu_i(\bar\theta_i)}(s_i(\bar\theta_i), \bar\theta_i, \{\emptyset\}) \geq U_i^{\mu_i(\bar\theta_i)}(s_i, \bar\theta_i, \{\emptyset\}) \qquad \text{for all } s_i \in S_i.$$

Hence $s_i(\bar\theta_i')$ is optimal at all $\mathcal{H} \in \mathscr{H}_i(s_i(\bar\theta_i'), \hat{s}_{-i}^{\bar\theta_i,\{\emptyset\}})$. Second, by 5. from Step 1, $s_i(\bar\theta_i')$ is also optimal for any other $\mathcal{H} \in \mathscr{H}_i(s_i(\bar\theta_i'))$ such that $\mathcal{H} \cap H_o^{\bar\theta_i' \leftarrow \bar\theta_i} \neq \emptyset$. Therefore, the desired $s_i^{\bar\theta_i' \leftarrow \bar\theta_i} \in Q_i^\infty(\bar\theta_i)$ can be constructed similarly as in Step 1.

Step 3. Let $k \in \mathbb{N}$. We claim that $s_i^{\theta_i' \leftarrow \theta_i} \in Q_i^k(\theta_i)$ for each $(i, \theta_i, \theta_i') \in N$ implies that $s_i^{\theta_i' \leftarrow \theta_i} \in Q_i^{k+1}(\theta_i)$ for each $(i, \theta_i, \theta_i') \in N$. For $(i, \bar\theta_i, \bar\theta_i') \in N$ let μ_i be a CPS such that

$$\mu_i((s_{-i}^{\bar\theta_i' \leftarrow \bar\theta_i \leftarrow \theta_{-i}}, \theta_{-i})|\{\emptyset\}) = \psi_i^{\bar\theta_i' \leftarrow \bar\theta_i}(\theta_{-i}) \qquad \text{for all } \theta_{-i} \in supp(\psi_i^{\bar\theta_i' \leftarrow \bar\theta_i})$$

and such that μ_i satisfies conditions 2. and 3. from Step 1. Then $\mu_i \in \Pi_i^k$ by Steps 1 and 2. Since for any $s_i \in S_i$ and any $\theta_{-i} \in supp(\psi_i^{\bar\theta_i' \leftarrow \bar\theta_i})$

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

the argument from 4. shows that $s_i^{\bar\theta_i' \leftarrow \bar\theta_i}$ maximizes $U_i^{\mu_i}(\cdot, \bar\theta_i, \mathcal{H})$ within $S_i(\mathcal{H})$ for any $\mathcal{H} \in \mathscr{H}_i(s_i(\bar\theta_i'), s_{-i}^{\bar\theta_i' \leftarrow \bar\theta_i \leftarrow \theta_{-i}}))$. Moreover, the argument from 5. applies without change. Hence $s_i^{\bar\theta_i' \leftarrow \bar\theta_i} \in r_i(\bar\theta_i, \mu_i)$ and so $s_i^{\bar\theta_i' \leftarrow \bar\theta_i} \in Q_i^{k+1}(\bar\theta_i)$.

Step 4. Therefore, $s_i^{\theta_i' \leftarrow \theta_i} \in Q_i^\infty(\theta_i)$ for all $(i, \theta_i, \theta_i') \in N$. Since β is unacceptable, $N \neq \emptyset$ and there is some $(i, \theta_i, \theta_i') \in N$. By definition, $f(\theta_i, \theta_{-i}^{\theta_i' \leftarrow \theta_i}) \neq f(\theta_i', \theta_{-i}^{\theta_i' \leftarrow \theta_i})$. Because Γ wr-implements f,

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

But since $s_i^{\theta_i' \leftarrow \theta_i}(\mathcal{H}) = s_i(\theta_i')(\mathcal{H})$ for all $\mathcal{H} \in \mathscr{H}_i$ containing a history in $H^{\theta_i' \leftarrow \theta_i}$, and in particular for all $\mathcal{H} \in \mathscr{H}_i$ containing a predecessor of $\zeta(s_i(\theta_i'), s_{-i}(\theta_{-i}^{\theta_i' \leftarrow \theta_i}))$, we must have $\zeta(s_i^{\theta_i' \leftarrow \theta_i}, s_{-i}(\theta_{-i}^{\theta_i' \leftarrow \theta_i})) = \zeta(s_i(\theta_i'), s_{-i}(\theta_{-i}^{\theta_i' \leftarrow \theta_i}))$. Contradiction. □

3.4 Ex-post Incentive Compatibility

Bergemann and Morris (2005) show that ex-post incentive compatibility (epIC) is necessary for partial robust implementation. Full robust implementation requires a stronger version of epIC called semi-strict epIC, which is implied by robust monotonicity (Bergemann and Morris, 2011, Lemma 1, Theorem 1). In this subsection, we show that (full) weakly rationalizable implementation implies a condition that is stronger than epIC but weaker than semi-strict epIC and that we call medium-strict epIC.

Definition 9 *Social choice function f is*

- *semi-strict epIC if for all $i \in \mathcal{I}$, $\theta_i, \theta_i' \in \Theta_i$ and $\theta_{-i} \in \Theta_{-i}$, if $f(\theta_i, \bar{\theta}_{-i}) \neq f(\theta_i', \bar{\theta}_{-i})$ for some $\bar{\theta}_{-i} \in \Theta_{-i}$ then*
$$u_i(f(\theta), \theta) > u_i(f(\theta_i', \theta_{-i}), \theta).$$
- *medium-strict epIC if for all $i \in \mathcal{I}$, $\theta_i, \theta_i' \in \Theta_i$ and $\theta_{-i} \in \Theta_{-i}$, if $f(\theta_i, \theta_{-i}) \neq f(\theta_i', \theta_{-i})$ then*
$$u_i(f(\theta), \theta) > u_i(f(\theta_i', \theta_{-i}), \theta).$$
- *epIC if for all $i \in \mathcal{I}$, $\theta_i, \theta_i' \in \Theta_i$ and all $\theta_{-i} \in \Theta_{-i}$*
$$u_i(f(\theta), \theta) \geq u_i(f(\theta_i', \theta_{-i}), \theta).$$

Proposition 1 revealed that dr-monotonicity is necessary for wr-implementation. The following proposition shows that medium-strict epIC, in turn, is necessary for dr-monotonicity. Together, these propositions establish that medium-strict epIC is the incentive-compatibility condition that is necessary for wr-implementation.

Proposition 2 *If f is dr-monotone then f is medium-strict epIC.*

Proof. Take $i \in \mathcal{I}$, $\theta_i, \theta_i' \in \Theta_i$ and $\theta_{-i} \in \Theta_{-i}$ and suppose that $f(\theta_i, \theta_{-i}) \neq f(\theta_i', \theta_{-i})$. Then the deception β such that $\beta_i(\theta_i) = \{\theta_i, \theta_i'\}$ and $\beta_j(\hat{\theta}_j) = \{\hat{\theta}_j\}$ for all $(j, \hat{\theta}_j) \neq (i, \theta_i)$ is unacceptable and hence, by hypothesis, d-refutable. The pair (θ_i', θ_i) is the only pair $(\hat{\theta}_j', \hat{\theta}_j)$ such that $j \in \mathcal{I}$, $\hat{\theta}_j' \in \beta_j(\hat{\theta}_j)$ and $\Theta_{-j}^{\hat{\theta}_j' \leftarrow \hat{\theta}_j} = \{\theta_{-j}' : f(\hat{\theta}_j, \theta_{-j}') \neq f(\hat{\theta}_j', \theta_{-j}')\} \neq \emptyset$. Moreover,

$\theta_{-i} \in \Theta_{-i}^{\theta_i' \leftarrow \theta_i}$. Hence, by the definition of d-refutability, for $\psi_i \in \Delta(\Theta_{-i})$ with $\psi_i\{\theta_{-i}\} = 1$ at least one of the following conditions holds:

$$\exists x \in Y_i(\theta_{-i}) : E_{\psi_i} u_i(x, \theta) > E_{\psi_i} u_i(f(\theta_i', \theta_{-i}), \theta) \tag{13}$$

$$\begin{aligned} \exists j \neq i, \theta_j \in supp(\psi_i), \psi_j^{BR} \in \Delta(\Theta_{-j}), y \in Y\ \forall \psi_j \in \Delta(\Theta_{-j})\ \exists x \in Y : \\ E_{\psi_j} u_j(x, \theta) > E_{\psi_j} u_j(y, \theta) \text{ and } E_{\psi_j^{BR}} u_j(x, \theta) \leq E_{\psi_j^{BR}} u_j(y, \theta) \end{aligned} \tag{14}$$

Condition (14) cannot hold, as it implies that for $\psi_j = \psi_j^{BR}$ there exists $x \in Y$ such that both $E_{\psi_j^{BR}} u_j(x, \theta) > E_{\psi_j^{BR}} u_j(y, \theta)$ and $E_{\psi_j^{BR}} u_j(x, \theta) \leq E_{\psi_j^{BR}} u_j(y, \theta)$. Consequently, (13) holds, and there exists $x \in Y_i(\theta_{-i})$ such that $u_i(x, \theta) > u_i(f(\theta_i', \theta_{-i}), \theta)$. Since $x \in Y_i(\theta_{-i})$ implies $u_i(f(\theta), \theta) \geq u_i(x, \theta)$,

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

follows. □

The following example presents a wr-implementable and therefore dr-monotone social choice function that is not semi-strict epIC. Consequently, dr-monotone social choice functions need to be medium-strict but not semi-strict epIC. The example also highlights the second of the differences between refutability and d-refutability mentioned at the beginning of Subsection 3.2.

Example 3.2 There are two agents $i \in \{1, 2\}$ with two payoff types each, $\Theta_i = \{\theta_i, \theta_i'\}$, and three pure outcomes, $X = \{x, y, z\}$. Player 1 prefers "not z" when she is of payoff type θ_1 and z when she is of payoff type θ_1':

$$\begin{aligned} u_1(x, \theta_1, \cdot) = u_1(y, \theta_1, \cdot) > u_1(z, \theta_1, \cdot) \\ u_1(z, \theta_1', \cdot) > u_1(x, \theta_1', \cdot) = u_1(y, \theta_1', \cdot) \end{aligned}$$

Type θ_2 of player 2 prefers x over y, and type θ_2' prefers y over x. Player 2's preference for z is not critical for the example, we assume 2 always prefers z the least:

$$\begin{aligned} u_2(x, \cdot, \theta_2) > u_2(y, \cdot, \theta_2) = u_2(z, \cdot, \theta_2) \\ u_2(y, \cdot, \theta_2') > u_2(x, \cdot, \theta_2') = u_2(z, \cdot, \theta_2') \end{aligned}$$

The social choice function f given in figure 5 is medium-strict epIC. Moreover, f is wr-implementable via the mechanism Γ and thus dr-monotone: truth-telling dominates lying for player 1, and conditionally (on making a decision) dominates lying for player 2. But

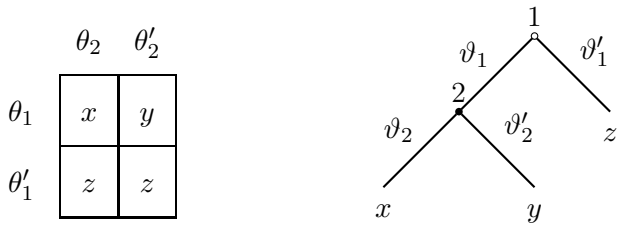

Figure 5: f (left) and mechanism Γ that wr-implements f (right)

f is not semi-strict epIC, as $f(\bar{\theta}_1, \theta_2) \neq f(\bar{\theta}_1, \theta_2')$ for some $\bar{\theta}_1 \in \Theta_1$, namely $\bar{\theta}_1 = \theta_1$, but $u_2(f(\theta_1', \theta_2), \theta_1', \theta_2) \not\succ u_2(f(\theta_1', \theta_2'), \theta_1', \theta_2)$. Consequently, f is neither robustly monotone nor robustly implementable by a static mechanism. Note that the unacceptable deception β such that $\beta_1(\hat{\theta}_1) = \{\hat{\theta}_1\}$ for all $\hat{\theta}_1 \in \Theta_1$, $\beta_2(\theta_2) = \Theta_2$ and $\beta_2(\theta_2) = \{\theta_2'\}$ is

- not refutable but...
  Proof: The triple $(2, \theta_2', \theta_2)$ is the only triple $(j, \hat{\theta}_j', \hat{\theta}_j)$ such that $j \in \mathcal{I}$, $\hat{\theta}_j \in \Theta_j$, $\hat{\theta}_j' \in \beta_j(\hat{\theta}_j)$ and $\Theta_{-j}^{\hat{\theta}_j' \leftarrow \hat{\theta}_j} \neq \emptyset$. But for 2, θ_2, $\theta_2' \in \beta_2(\theta_2)$ and θ_1' and $\psi_2 = \delta(\theta_1')$, there does not exist a $w \in Y_2(\theta_1') = \{z\}$ such that $E_{\psi_2} u_2(w, \theta) = u_2(w, \theta_1', \theta_2) > u_2(z, \theta_1', \theta_2) = E_{\psi_2} u_2(f(\theta'), \theta)$.

- ... d-refutable.
  Proof: For 2, θ_2 and $\theta_2' \in \beta_2(\theta_2)$, $\Theta_1^{\theta_2' \leftarrow \theta_2} = \{\theta_1\} \neq \emptyset$ and for all ψ_2 such that $\psi_2\{\theta_1\} = 1$, x is in $Y_2(\theta_1)$ and satisfies $E_{\psi_2} u_2(x, \theta) = u_2(x, \theta) > u_2(y, \theta) = E_{\psi_2} u_2(f(\theta_1, \theta_2'), \theta)$.

The unacceptable deception β thus illustrates the second of the reasons (listed at the beginning of Subsection 3.2) that d-refutability is weaker than refutability: For 2, θ_2 and $\theta_2' \in \beta_2(\theta_2)$ we have $\Theta_1^{\theta_2' \leftarrow \theta_2} \neq \emptyset$. Moreover, for all $\hat{\theta}_1 \in \Theta_1^{\theta_2' \leftarrow \theta_2}$ — but *not* for all $\hat{\theta}_1 \in \Theta_1$ — and all $\psi_2 \in \Delta(\Theta_1)$ with $\psi_2(\beta_1^{-1}(\hat{\theta}_1)) = 1$, condition (2) of Definition 6 holds.

4 Sufficient Conditions for WR-Implementation

For the purpose of deriving sufficient conditions for wr-implementation, it is convenient to focus on semi-strict epIC social choice functions. In addition, we assume the following NTI condition, taken from Bergemann and Morris (2011). The conditional NTI condition is a mild restriction on the players' preferences.

Definition 10 (Conditional NTI.) *The conditional no total indifference (NTI) property is met if for all $i \in \mathcal{I}$, $\theta_i \in \Theta_i$, $\theta_{-i}' \in \Theta_{-i}$ and $\psi_i \in \Delta(\Theta_{-i})$, there exist $y, y' \in Y_i(\theta_{-i}')$ such that*

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

Using infinite mechanisms, we are able to establish dr-monotonicity as sufficient for wr-implementation. Our mechanism 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 is a property that our mechanism shares with Bergemann and Morris' (2011) mechanism and other static integer and modulo game forms. This property has been criticized by Jackson (1992) among 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.

Proposition 3 *Suppose the conditional NTI property is satisfied. If f is semi-strict epIC and dr-monotone, then f is wr-implementable.*

Proof. First, some preliminaries. For $i \in \mathcal{I}$ and $\theta_{-i} \in \Theta_{-i}$, 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 $\{y \in Y : u_i(y, \theta_i'', \theta_{-i}) \leq u_i(f(\theta_i'', \theta_{-i}), \theta_i'', \theta_{-i})\}$, $\theta_i'' \in \Theta_i$. As such, $Y_i(\theta_{-i})$ is convex and has finitely many extreme points $y_{1,\theta_{-i}}, y_{2,\theta_{-i}}, \ldots, y_{m,\theta_{-i}}$. Let $\bar{y}_{\theta_{-i}}$ be the convex combination that puts weight $\frac{1}{m}$ on each extreme point. Then by the conditional NTI property this "interior" lottery $\bar{y}_{\theta_{-i}} \in Y_i(\theta_{-i})$ is such that

6. for every $\theta_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i})$ there is a $y \in Y_i(\theta_{-i})$ such that $E_{\psi_i} u_i(y, \theta) > E_{\psi_i} u_i(\bar{y}_{\theta_{-i}}, \theta)$ (compare Bergemann and Morris, 2011, p. 270) and ...
   Proof: Let $\theta_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i})$, then by the conditional NTI property 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, $E_{\psi_i} u_i(\bar{y}_{\theta_{-i}} + \eta(y - y'), \theta) > E_{\psi_i} u_i(\bar{y}_{\theta_{-i}}, \theta)$.

7. ... for every $\theta_i \in \Theta_i$, $u_i(\bar{y}_{\theta_{-i}}, \theta) < u_i(f(\theta), \theta)$.
   Proof: Let $\theta_i \in \Theta_i$, then by the conditional NTI property 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 $\bar{k}$. This implies the claim.

For each unacceptable (and hence by hypothesis 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 $\Theta_{-i}^{\theta_i' \leftarrow \theta_i} \neq \emptyset$ and such that for all

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

$$\exists x \in Y_i(\theta'_{-i}) : E_{\psi_i} u_i(x, \theta) > E_{\psi_i} u_i(f(\theta'), \theta) \tag{15}$$

or

$$\begin{aligned} &\exists j \neq i, \theta_j \in supp(\psi_i), \psi_j^{BR} \in \Delta(\Theta_{-j}), y \in Y \ \forall \psi_j \in \Delta(\Theta_{-j}) \ \exists x \in Y : \\ &\qquad E_{\psi_j} u_j(x, \theta) > E_{\psi_j} u_j(y, \theta) \text{ and } E_{\psi_j^{BR}} u_j(x, \theta'_j, \theta_{-j}) \leq E_{\psi_j^{BR}} u_j(y, \theta'_j, \theta_{-j}). \end{aligned} \tag{16}$$

To streamline notation, let

$$\begin{aligned} \mathbb{B}^{(15)} = \Big\{ &(\beta, \theta'_{-i}, \psi_i) : \beta \text{ is an unacceptable deception, } i = i(\beta), \theta'_{-i} \in \Theta_{-i}^{\theta'_i(\beta) \leftarrow \theta_i(\beta)}, \\ &\psi_i \in \Delta(\Theta_{-i}), \psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1 \text{ and } (15) \Big\} \end{aligned}$$

and

$$\begin{aligned} \mathbb{B}^{(16)} = \Big\{ &(\beta, \theta'_{-i}, \psi_i) : \beta \text{ is an unacceptable deception, } i = i(\beta), \theta'_{-i} \in \Theta_{-i}^{\theta'_i(\beta) \leftarrow \theta_i(\beta)}, \\ &\psi_i \in \Delta(\Theta_{-i}), \psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1 \text{ and } \neg(15) \Big\}. \end{aligned}$$

If $(\beta, \theta'_{-i}, \psi_i) \in \mathbb{B}^{(15)}$ let $x^{\beta, \theta'_{-i}, \psi_i} \in Y_i(\theta'_{-i})$ be such that (15) holds. Similarly, if $(\beta, \theta'_{-i}, \psi_i) \in \mathbb{B}^{(16)}$, let $j^{\beta, \theta'_{-i}, \psi_i}$, $\theta_j^{\beta, \theta'_{-i}, \psi_i}$, $\psi_j^{BR, \beta, \theta'_{-i}, \psi_i}$ and $y^{\beta, \theta'_{-i}, \psi_i}$ be such that for every ψ_j (where $j = j^{\beta, \theta'_{-i}, \psi_i}$) there is an $x \in Y$, let's denote it by $x^{\beta, \theta'_{-i}, \psi_i}(\psi_j)$, such that (16) holds.

We construct a mechanism $\Gamma = \langle H, (\mathscr{H}_i)_{i \in \mathcal{I}}, P, C \rangle$ that is an augmented direct mechanism with two stages. For each $(\beta, \theta'_{-i}, \psi_i) \in \mathbb{B}^{(16)}$ let $m_i^{\beta, \theta'_{-i}, \psi_i}$ be some action (or "message") that is distinct from all other actions. In the first stage of Γ we let the agents make strategically simultaneous announcements, with each agent being able to announce a payoff type or one of the messages $m_i^{\beta, \theta'_{-i}, \psi_i}$. The set of i's first-stage actions is

$$A_i^1 = \Theta_i \cup \bigcup_{(\beta, \theta'_{-i}, \psi_i) \in \mathbb{B}^{(16)}} \{m_i^{\beta, \theta'_{-i}, \psi_i}\},$$

and the set of first-stage histories $H^1 = \{h : h \preceq a \text{ for some } a \in \prod_{i \in \mathcal{I}} A_i^1\}$. If $h = (a_1, \ldots, a_{i-1})$ then $P(h) = i$, and $\mathcal{H}_i^1 = \{(a_1, \ldots, a_{i-1}) \in H^1\} \in \mathscr{H}_i$. We complete the description of Γ's set of histories by defining a (potentially empty) set of histories H^a for each profile of first-stage actions $a = (a_1, \ldots, a_I) \in H^1$ and letting

$$H = H^1 \cup \bigcup_{a \in H^1} H^a.$$

Let d_i^{BR} and $d_i^{\psi_i, k_i}$ (where $i \in \mathcal{I}$, $\psi_i \in \Delta(\Theta_{-i})$ and $k_i \in \{2, 3, \ldots\}$) be some actions that are distinct from all other actions and distinguish the following cases.

Case 1. Suppose that the first-stage action profile is $\theta' \in \Theta$. If

$$\mathcal{I}(\theta') = \left\{ i \in \mathcal{I} : \left(\exists \beta, \psi_i \right) \left((\beta, \theta'_{-i}, \psi_i) \in \mathbb{B}^{(15)} \text{ and } \theta'_i = \theta'_i(\beta) \right) \right\}$$

is empty, then there is no second stage. We let $H^{\theta'} = \emptyset$ and assign the outcome $C(\theta') = f(\theta')$ to the terminal history θ'. If on the other hand $\mathcal{I}(\theta') \neq \emptyset$ then

$$H^{\theta'} = \Big\{ h : \Big(\exists (d_i)_{i \in \mathcal{I}(\theta')} \Big) \Big(h \preceq (\theta', (d_i)) \text{ and } \forall i \in \mathcal{I}(\theta'), \\ d_i \in \left\{ d_i^{BR} \right\} \cup \left\{ (z_i, k_i) \in Y_i(\theta'_{-i}) \times \{1, 2, 3, \ldots\} \right\} \Big) \Big\}.$$

All agents $i \in \mathcal{I}(\theta')$ make their second-stage decisions strategically simultaneously, with each $i \in \mathcal{I}(\theta')$ observing the first-stage history θ' before choosing her second-stage action d_i. Finally, let

$$C(\theta', (d_i)_{i \in \mathcal{I}(\theta')}) = \frac{1}{\#\mathcal{I}(\theta')} \sum_{i \in \mathcal{I}(\theta')} C_i(\theta', d_i),$$

where

$$C_i(\theta', d_i) = \begin{cases} f(\theta') & \text{if } d_i = d_i^{BR} \\ \frac{1}{k_i} \bar{y}_{\theta'_{-i}} + \left(1 - \frac{1}{k_i}\right) z_i & \text{if } d_i = (z_i, k_i) \end{cases}.$$

Case 2. If exactly one agent $i \in \mathcal{I}$ announces a first-stage message $a_i \notin \Theta_i$, let β, θ'_{-i}, ψ_i be such that $a_i = m_i^{\beta, \theta'_{-i}, \psi_i}$. Then (16) holds for i, $\theta_i(\beta)$ and $\theta'_i(\beta)$ and θ'_{-i}, ψ_i. If $a_{-i} \neq \theta'_{-i}$ then a will be a terminal history. Let $H^a = \emptyset$ and $C(a) = \bar{y}_{a_{-i}}$. If $a_{-i} = \theta'_{-i}$ then after learning the first-stage messages a, the agents i and $j = j^{\beta, \theta'_{-i}, \psi_i}$ make strategically simultaneous second-stage choices, and

$$H^a = \Big\{ h : h \preceq (a, d_j, d_i) \text{ for some } d_j \in \left\{ d_j^{BR} \right\} \cup \bigcup_{\psi_j \in \Delta(\Theta_{-j}), k_j \in \{2,3,\ldots\}} \left\{ d_j^{\psi_j, k_j} \right\} \\ \text{and some } d_i = (z_i, k_i) \in Y_i(a_{-i}) \times \{2, 3, \ldots\} \Big\}.$$

Let

$$C_i(a, z_i, k_i) = \left(\frac{1}{2} + \frac{1}{k_i} \right) \bar{y}_{a_{-i}} + \left(\frac{1}{2} - \frac{1}{k_i} \right) z_i,$$

$C_j(a, d_j^{BR}) = y^{\beta, \theta'_{-i}, \psi_i}$ and

$$C_j(a, d_j^{\psi_j, k_j}) = \frac{1}{k_j} y^{\beta, \theta'_{-i}, \psi_i} + \left(1 - \frac{1}{k_j} \right) x^{\beta, \theta'_{-i}, \psi_i}(\psi_j). \tag{17}$$

By 7. we can, for each $\theta_i \in \Theta_i$, let $\varepsilon(\theta_i) \in (0,1)$ be such that

$$\frac{1}{2}(1-\varepsilon(\theta_i))u_i(\bar{y}_{a_{-i}},\theta_i,a_{-i}) + \varepsilon(\theta_i)u_i(y^{\beta,a_{-i},\psi_i},\theta_i,a_{-i}) \leq \left(\frac{1}{2}+\frac{1}{2}\varepsilon(\theta_i)\right)u_i(f(\theta_i,a_{-i}),\theta_i,a_{-i}). \tag{18}$$

Let $\varepsilon = \min_{\theta_i\in\Theta_i}\varepsilon(\theta_i)$ and $C(a,d_j,d_i) = (1-\varepsilon)C_i(a,d_i) + \varepsilon C_j(a,d_j)$.

Case 3. If $a_i \notin \Theta_i$ for two or more $i \in \mathcal{I}$, then let a be terminal, $H^a = \emptyset$ and $C(a) \in Y$.

The set of histories H, the set of i's information sets[17]

$$\mathscr{H}_i = \{\mathcal{H}_i^1\} \cup \bigcup_{\theta'\in\Theta \text{ s.t. } i\in\mathcal{I}(\theta')} \{(\theta',(d_j)_{j<i,j\in\mathcal{I}(\theta')}) \in H\} \cup \bigcup_{(\beta,\theta'_{-i},\psi_i)\in\mathbb{B}^{(16)}} \{(m_i^{\beta,\theta'_{-i},\psi_i},\theta'_{-i},d_j)\in H\}$$
$$\cup \bigcup_{l\neq i,\,(\beta,\theta'_{-l},\psi_l)\in\mathbb{B}^{(16)} \text{ s.t. } i=j^{\beta,\theta'_{-l},\psi_l}} \{(m_l^{\beta,\theta'_{-l},\psi_l},\theta'_{-l})\}$$

for $i \in \mathcal{I}$, the outcome function C and the implied player function P fully describe the mechanism Γ. We now prove that Γ wr-implements f.

Step 1. If $(s_i,\theta_i) \in W_i^1$ then $s_i(\mathcal{H}_i^1) \in \Theta_i$.

Proof: Suppose that $s_i(\mathcal{H}_i^1) = m_i^{\beta,\theta'_{-i},\psi_i} \notin \Theta_i$ for some β, θ'_{-i} and ψ_i, and that s_i is sequentially rational for $\theta_i \in \Theta_i$. Then there exists a CPS μ_i such that for all $\mathcal{H} \in \mathscr{H}_i(s_i)$, s_i maximizes $U_i^{\mu_i}(\cdot,\theta_i,\mathcal{H})$. In particular, s_i must be optimal in case $-i$ announces θ'_{-i} in the first stage. That is, for $\mathcal{H} = \{(m_i^{\beta,\theta'_{-i},\psi_i},\theta'_{-i},d_j) \in H\}$, $U_i^{\mu_i}(s_i,\theta_i,\mathcal{H}) \geq U_i^{\mu_i}(s_i',\theta_i,\mathcal{H})$ for all $s_i' \in S_i(\mathcal{H})$. Let $(z_i,k_i) \in Y_i(\theta'_{-i}) \times \{2,3,\ldots\}$ be such that $s_i(\mathcal{H}) = (z_i,k_i)$. Recall that for all d_j,

$$C(m_i^{\beta,\theta'_{-i},\psi_i},\theta'_{-i},d_j,z_i,k_i) = \varepsilon C_j(m_i^{\beta,\theta'_{-i},\psi_i},\theta'_{-i},d_j) + (1-\varepsilon)\left(\left(\frac{1}{2}+\frac{1}{k_i}\right)\bar{y}_{\theta'_{-i}} + \left(\frac{1}{2}-\frac{1}{k_i}\right)z_i\right).$$

Since by 6. we have

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

we must have $E_{marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H})}u_i(z_i,\theta) > E_{marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H})}u_i(\bar{y}_{\theta'_{-i}},\theta)$ and $k_i \geq 3$. But then choosing (z_i,k_i+1) at $\mathcal{H}$ makes θ_i better off at $\mathcal{H}$ than s_i. Contradiction.

Step 2. If $(s_i,\bar{\theta}_i) \in W_i^1$ and $\mathcal{H}_i^2 = \{(\theta',(d_j)_{j<i,j\in\mathcal{I}(\theta')}) \in H\} \in \mathscr{H}_i(s_i)\backslash\{\mathcal{H}_i^1\}$ for some $\theta' \in \Theta$, then $s_i(\mathcal{H}_i^2) = d_i^{BR}$.

[17] In this proof, we use concise notation for information sets. E.g., given $\theta' \in \Theta$, $\{(\theta',(d_j)_{j<i,j\in\mathcal{I}(\theta')}) \in H\}$ denotes $\{(\theta,(d_j)_{j<i,j\in\mathcal{I}(\theta')}) \in H : \theta = \theta'\}$.

Proof: By Step 1, if s_i is rational for $\bar{\theta}_i$ then $s_i(\mathcal{H}_i^1) = \theta_i'$ for some $\theta_i' \in \Theta_i$. If for $\theta_{-i}' \in \Theta_{-i}$, $\mathcal{H}_i^2 = \{(\theta', (d_j)_{j<i, j\in\mathcal{I}(\theta')}) \in H\}$ is a second-stage information set in $\mathscr{H}_i(s_i)$ then $i \in \mathcal{I}(\theta')$. Suppose that $s_i(\mathcal{H}_i^2) \neq d_i^{BR}$ then $s_i(\mathcal{H}_i^2) = (z_i, k_i)$ for some $(z_i, k_i) \in Y_i(\theta_{-i}') \times \{1, 2, 3, \ldots\}$. An argument similar to that of Step 1 shows that s_i cannot be sequentially rational as $\bar{\theta}_i$ prefers to play $(z_i, k_i + 1)$ instead of (z_i, k_i) at $\mathcal{H}_i^2$.

Step 3. If $(\beta, \theta_{-i}', \psi_i) \in \mathbb{B}^{(16)}$, $l = j^{\beta,\theta_{-i}',\psi_i}$, $\bar{\theta}_l = \theta_j^{\beta,\theta_{-i}',\psi_i}$ and $s_l \in Q_l^1(\bar{\theta}_l)$, then $s_l(\mathcal{H}_l^1) \neq \theta_l'$.

Proof: Suppose that $s_l(\mathcal{H}_l^1) = \theta_l'$, then s_l admits l's second stage information set $\mathcal{H} = \{(m_i^{\beta,\theta_{-i}',\psi_i}, \theta_{-i}')\}$. Suppose further that $(s_l, \bar{\theta}_l) \in \rho_l(\mu_l)$ for some CPS μ_l and let $\psi_l \in \Delta(\Theta_{-l})$, $k_l \geq 1$ be such that $C_l(m_i^{\beta,\theta_{-i}',\psi_i}, \theta_{-i}', s_l(\mathcal{H})) = \frac{1}{k_l} y^{\beta,\theta_{-i}',\psi_i} + (1 - \frac{1}{k_l}) x^{\beta,\theta_{-i}',\psi_i}(\psi_l)$. By (16), we must have

$$E_{marg_{\Theta_{-l}}\mu_l(\cdot|\mathcal{H})} u_l(x^{\beta,\theta_{-i}',\psi_i}(\psi_l), \bar{\theta}_l, \theta_{-l}) > E_{marg_{\Theta_{-l}}\mu_l(\cdot|\mathcal{H})} u_l(y^{\beta,\theta_{-i}',\psi_i}, \bar{\theta}_l, \theta_{-l}).$$

Therefore, playing $d_l^{\psi_l, k_l+1}$ instead of $s_l(\mathcal{H})$ gives $\bar{\theta}_l$ a higher expected utility at $\mathcal{H}$. Contradiction, hence $(s_l, \bar{\theta}_l) \notin \rho_l(\mu_l)$ and $s_l \notin Q_l^1(\bar{\theta}_l)$.

Step 4. If $s_i \in S_i$ satisfies $s_i(\mathcal{H}_i^1) = \bar{\theta}_i$ and $s_i(\mathcal{H}_i^2) = d_i^{BR}$ for all $\mathcal{H}_i^2 \in \mathscr{H}_i(s_i) \backslash \{\mathcal{H}_i^1\}$, then $s_i \in Q_i^\infty(\bar{\theta}_i)$.

Proof: Let

$$S_i(\theta_i) = \left\{ s_i \in S_i : s_i(\mathcal{H}_i^1) = \theta_i \text{ and } s_i(\mathcal{H}_i^2) = d_i^{BR} \text{ for all } \mathcal{H}_i^2 \in \mathscr{H}_i(s_i) \backslash \{\mathcal{H}_i^1\} \right\}$$

for all $i \in \mathcal{I}$, $\theta_i \in \Theta_i$. Let $\bar{s}_i \in S_i(\bar{\theta}_i)$ and let μ_i be a CPS such that 1) for all $\mathcal{H} \in \bar{\mathscr{H}}_i(\{\bar{s}_i\} \times \bigcup_{\theta_{-i}\in\Theta_{-i}} S_{-i}(\theta_{-i}))$ there are $\bar{\theta}_{-i} \in \Theta_{-i}$ and $s_{-i}^{\mathcal{H}} \in S_{-i}(\bar{\theta}_{-i})$ such that $\mu_i((s_{-i}^{\mathcal{H}}, \bar{\theta}_{-i})|\mathcal{H}) = 1$ and 2) for all $\mathcal{H} = \{(\bar{\theta}_i, m_l^{\beta,(\bar{\theta}_i,\theta_{-i,l}'),\psi_l}, \theta_{-i,l}') \in H\} \in \mathscr{H}_i$ there is some $s_{-i}^{\mathcal{H}} \in S_{-i}(\mathcal{H})$ such that $\mu_i((s_{-i}^{\mathcal{H}}, \theta_{-i})|\mathcal{H}) = \psi_i^{BR,\beta,(\bar{\theta}_i,\theta_{-i,l}'),\psi_l}(\theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$. We claim that $\bar{s}_i \in r_i(\bar{\theta}_i, \mu_i)$.

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

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

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

- for any $s_i \in S_i(\mathcal{H})$ such that $s_i(\mathcal{H}_i^1) \in \Theta_i$

$$\zeta(s_i, s_{-i}^{\mathcal{H}}) \in \left\{ \alpha \left(\frac{1}{k_i} \bar{y}_{\bar{\theta}_{-i}} + \left(1 - \frac{1}{k_i}\right) z_i \right) + (1 - \alpha) f(\theta_i', \bar{\theta}_{-i}) : \right.$$
$$\left. \theta_i' \in \Theta_i, z_i \in Y_i(\bar{\theta}_{-i}), k_i \geq 1, \alpha \in [0, 1] \right\}$$

and $\bar{\theta}_i$ prefers $f(\bar{\theta})$ over $f(\theta'_i, \bar{\theta}_{-i})$ (the social choice function f is epIC) and over any element of $Y_i(\bar{\theta}_{-i})$ (by construction of the reward set).

- for any $s_i \in S_i(\mathcal{H})$ such that $s_i(\mathcal{H}_i^1) \notin \Theta_i$

$$\zeta(s_i, s_{-i}^{\mathcal{H}}) \in \left\{\bar{y}_{\bar{\theta}_{-i}}\right\} \cup \left\{(1-\varepsilon)\left(\left(\frac{1}{2}+\frac{1}{k_i}\right)\bar{y}_{\bar{\theta}_{-i}} + \left(\frac{1}{2}-\frac{1}{k_i}\right)z_i\right) + \varepsilon y^{\beta,\bar{\theta}_{-i},\psi_i} : \right.$$
$$\left. z_i \in Y_i(\bar{\theta}_{-i}), k_i \geq 2, (\beta, \psi_i) \text{ s.t. } (\beta, \bar{\theta}_{-i}, \psi_i) \in \mathbb{B}^{(16)}\right\}$$

for some ε such that (18) holds and $\bar{\theta}_i$ prefers $f(\bar{\theta})$ over any element of $Y_i(\bar{\theta}_{-i})$ (by construction of the reward set).

Next, consider $\mathcal{H} = \{a\} \in \mathscr{H}_i$ where $a = (\bar{\theta}_i, m_l^{\beta,(\bar{\theta}_i,\theta'_{-i,l}),\psi_l}, \theta'_{-i,l})$. Condition (16) holds for $l = i(\beta)$, $\theta_l(\beta)$ and $\theta'_l(\beta)$ and $(\bar{\theta}_i, \theta'_{-i,l})$, ψ_l. Moreover, $i = j^{\beta,(\bar{\theta}_i,\theta'_{-i,l}),\psi_l}$. We have

$$U_i^{\mu_i}(\bar{s}_i, \bar{\theta}_i, \mathcal{H}) \geq U_i^{\mu_i}(s_i, \bar{\theta}_i, \mathcal{H})$$

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

$$\zeta(\bar{s}_i, s_{-i}^{\mathcal{H}}) = (1-\varepsilon)C_l(a, s_l^{\mathcal{H}}(\mathcal{H})) + \varepsilon y^{\beta,(\bar{\theta}_i,\theta'_{-i,l}),\psi_l}$$

and

$$\zeta(s_i, s_{-i}^{\mathcal{H}}) = (1-\varepsilon)C_l(a, s_l^{\mathcal{H}}(\mathcal{H})) + \varepsilon\left(\frac{1}{k_i}y^{\beta,(\bar{\theta}_i,\theta'_{-i,l}),\psi_l} + \left(1-\frac{1}{k_i}\right)x^{\beta,(\bar{\theta}_i,\theta'_{-i,l}),\psi_l}(\psi_i)\right)$$

for some $\varepsilon \in (0,1)$, $k_i \geq 2$ and $\psi_i \in \Delta(\Theta_{-i})$, and by the definition of μ_i and by (16),

$$E_{marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H})}u_i(y^{\beta,(\bar{\theta}_i,\theta'_{-i,l}),\psi_l}, \bar{\theta}_i, \theta_{-i}) \geq E_{marg_{\Theta_{-i}}\mu_i(\cdot|\mathcal{H})}u_i(x^{\beta,(\bar{\theta}_i,\theta'_{-i,l}),\psi_l}(\psi_i), \bar{\theta}_i, \theta_{-i})$$

for all $\psi_i \in \Delta(\Theta_{-i})$.

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 induction, $\bar{s}_i \in Q_i^\infty(\bar{\theta}_i)$.

Step 5. If $\bar{s}_i \in Q_i^\infty(\bar{\theta}_i)$ and $\bar{s}_i(\mathcal{H}_i^1) = \theta'_i$, then $\Theta_{-i}^{\theta'_i \leftarrow \bar{\theta}_i} = \emptyset$.

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

$$\beta_i(\theta_i) = \{\theta'_i \in \Theta_i : \exists s_i \in Q_i^\infty(\theta_i), s_i(\mathcal{H}_i^1) = \theta'_i\}$$

(Step 4 ensures that β is indeed a deception). Suppose that β is unacceptable, and hence by hypothesis d-refutable. Then for $i = i(\beta)$, $\theta_i = \theta_i(\beta)$ and $\theta'_i = \theta'_i(\beta)$ and for each $\theta'_{-i} \in \Theta_{-i}^{\theta'_i \leftarrow \theta_i}$ and $\psi_i \in \Delta(\Theta_{-i})$ with $\psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1$, (15) or (16) holds. In fact, we know that (15) must

be true: Suppose (15) is false for θ'_{-i}, ψ_i. Then $(\beta, \theta'_{-i}, \psi_i) \in \mathbb{B}^{(16)}$. Let $l = j^{\beta,\theta'_{-i},\psi_i}$ and $\theta_l = \theta_j^{\beta,\theta'_{-i},\psi_i} \in supp(\psi_i)$. If $s_l \in Q_l^\infty(\theta_l)$ then $s_l \in Q_l^1(\theta_l)$ and $s_l(\mathcal{H}_l^1) \neq \theta'_l$ by Step 3. Hence $\theta'_l \notin \beta_l(\theta_l)$ by the definition of β, and $\theta_l \notin supp(\psi_i)$. Contradiction. Moreover, since $\Theta_{-i}^{\theta'_i \leftarrow \theta_i} \neq \emptyset$ and f is semi-strict epIC, $f(\theta'_i, \theta_{-i}) \neq f(\theta)$ for all $\theta_{-i} \in \Theta_{-i}$ and hence $\Theta_{-i}^{\theta'_i \leftarrow \theta_i} = \Theta_{-i}$.

Hence for i, θ_i, θ'_i and for all $\theta'_{-i} \in \Theta_{-i}$ and $\psi_i \in \Delta(\Theta_{-i})$ with $\psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1$, (15) is true. Let $s_i \in Q_i^\infty(\theta_i)$ be such that $s_i(\mathcal{H}_i^1) = \theta'_i$. Since $s_i \in Q_i^\infty(\theta_i)$ there must be some $\mu_i \in \Pi_i^\infty$ against which s_i is a sequential best response for θ_i. By Step 1, $(s_{-i}, \theta_{-i}) \in W_{-i}^\infty$ implies $(s_j(\mathcal{H}_j^1))_{j \neq i} \in \Theta_{-i} = \Theta_{-i}^{\theta'_i \leftarrow \theta_i}$. Therefore, there exists $\theta'_{-i} \in \Theta_{-i}$ such that $\mu_i(\bar{S}_{-i}(\theta'_{-i}) \times \Theta_{-i} | \mathcal{H}_i^1) > 0$, where

$$\bar{S}_{-i}(\theta'_{-i}) = \left\{ s_{-i} \in S_{-i} : \bar{s}_j(\mathcal{H}_j^1) = \theta'_j \text{ for all } j \neq i \right\}.$$

Let

$$\psi_i(\theta_{-i}) = \frac{\mu_i(\bar{S}_{-i}(\theta'_{-i}) \times \{\theta_{-i}\} | \mathcal{H}_i^1)}{\mu_i(\bar{S}_{-i}(\theta'_{-i}) \times \Theta_{-i} | \mathcal{H}_i^1)} \quad \text{for all } \theta_{-i} \in \Theta_{-i}.$$

Then $\psi_i(\beta_{-i}^{-1}(\theta'_{-i})) = 1$ and $(\beta, \theta'_{-i}, \psi_i) \in \mathbb{B}^{(15)}$ and $i \in \mathcal{I}(\theta')$. Therefore i has an information set $\mathcal{H}_i^2 = \{(\theta', (d_j)_{j<i, j \in \mathcal{I}(\theta')}) \in H\}$. Note that $\psi_i = marg_{\Theta_{-i}} \mu_i(\cdot | \mathcal{H}_i^2)$. By (15) for θ'_{-i} and ψ_i

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

and there is k_i large enough so that playing $(x^{\beta,\theta'_{-i},\psi_i}, k_i)$ at $\mathcal{H}_i^2$ provides θ_i with a strictly higher expected utility than playing $s_i(\mathcal{H}_i^2)$ (which by Step 2 equals d_i^{BR}). Contradiction to s_i being sequentially rational for θ_i with respect to μ_i. Hence β must be acceptable and Step 5's claim be true.

Step 6. By Steps 1,2 and 5, if the profile $s \in S$ is weakly rationalizable for $\theta \in \Theta$ then $C(\zeta(s)) = f(\theta)$. The proof of Step 4 implies that Γ is pepWB. Hence Γ wr-implements f. $\square$

This completes the proof of Proposition 3. Note that Steps 1 and 2 of the proof imply that the iterated elimination of never-best sequential best responses converges in finitely many rounds, so that there is a $k \in \mathbb{N}$ such that $W^{k'} = W^\infty$ for all $k' \geq k$.

A Mechanisms

This appendix formally defines a mechanism. We first recall the definition of an extensive game form (see e.g. Kuhn (1953); our notation is close to that of Osborne and Rubinstein (1994)).

Definition 11 *An extensive game form is a tuple* $\Gamma = \langle H, (\mathscr{H}_i)_{i \in \mathcal{I}}, P, C \rangle$ *such that*

- *H is a nonempty set of finite sequences with codomain A (where A is a nonempty metrizable topological space of actions) such that with h every initial subsequence of h is in H.*[18] *We let $A(h) = \{a \in A : (h,a) \in H\}$ for $h \in H$, let $T = \{h \in H : A(h) = \emptyset\}$ 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 $h \in H$.*

- *$P : H \backslash T \to \mathcal{I}$.*

- *for each $i \in \mathcal{I}$, $\mathscr{H}_i$ is a partition of $\{h \in H \backslash T : P(h) = i\}$ such that*[19]

  - *for all $\mathcal{H} \in \mathscr{H}_i$ and all $h, h' \in \mathcal{H}$, $A(h) = A(h')$.*
  - *for all $\mathcal{H} \in \mathscr{H}_i$ and all $h, h' \in H$, if $h \in \mathcal{H} \in \mathscr{H}_i$, $h' \preceq h$ and $h' \neq h$ then $h' \notin \mathcal{H}$.*

- *$C : T \to Y$.*

Note that we allow for infinitely many actions at any history but not for infinitely many time periods (histories are finite sequences), and that $\preceq$ partially orders H. We define a binary relation $\preceq$ on $\mathscr{H}_i$ by $\mathcal{H}' \preceq \mathcal{H}$ if there are $h' \in \mathcal{H}'$ and $h \in \mathcal{H}$ such that $h' \preceq h$, and extend this relation to $\bar{\mathscr{H}}_i = \mathscr{H}_i \cup \{\{\emptyset\}\}$ (if necessary) by letting $\{\emptyset\} \preceq \mathcal{H}$ for all $\mathcal{H} \in \bar{\mathscr{H}}_i$. A strategy for player i in an extensive game form Γ is a function $s_i : \mathscr{H}_i \to A$ such that for all $\mathcal{H} \in \mathscr{H}_i$, there is an $h \in \mathcal{H}$ such that $s_i(\mathcal{H}) \in A(h)$. The set of player i's strategies admitting information set $\mathcal{H} \in \mathscr{H}_j$, $j \in \mathcal{I}$, is defined as $S_i(\mathcal{H}) = \{s_i \in S_i : \exists s_{-i} \in S_{-i} \exists h \in \mathcal{H}, h \preceq \zeta(s)\}$.

To ensure that our definition of a Bayesian agent (made in Subsection 2.2) is sensible, we restrict attention to extensive game forms with perfect recall and no trivial decision nodes. In order to define perfect recall, we let $H(s_i) = \left\{h \in H : \exists s_{-i} \in S_{-i}, h \preceq \zeta(s)\right\}$ denote the set of histories admitted by $s_i \in S_i$.

Definition 12 *A mechanism is an extensive game form $\Gamma = \langle H, (\mathscr{H}_i)_{i \in \mathcal{I}}, P, C \rangle$ such that*

- (perfect recall) *for all $i \in \mathcal{I}$, $s_i \in S_i$ and $\mathcal{H} \in \mathscr{H}_i$, if $\mathcal{H} \cap H(s_i) \neq \emptyset$ then $\mathcal{H} \subseteq H(s_i)$.*

- (no trivial decisions) *for all $(h,a) \in H$ there exists an action $a' \neq a$ such that $(h,a') \in H$.*

A mechanism is *finite* if H is finite. A mechanism is *static* if each agent has exactly one information set, if at any two non-terminal histories of equal length the same player is active, and if all terminal histories have the same length.

[18] Let A be a nonempty set. A finite sequence h of length $n \in \mathbb{N}$ with codomain A is a function $h : \{1, \ldots, n\} \to A$ (where $\{1, \ldots, n\}$ denotes $\emptyset$ if $n = 0$). A finite sequence $g : \{1, \ldots, k\} \to A$ is an initial subsequence of the finite sequence $h : \{1, \ldots, n\} \to A$ if $k \leq n$ and $g_l = h_l$ for all $l \in \{1, \ldots, k\}$. Note that $\emptyset$ (the unique finite sequence mapping $\{1, \ldots, 0\}$ to A) is an initial subsequence of every finite sequence with codomain A. For $h : \{1, \ldots, n\} \to A$ and $a \in A$, (h,a) denotes the finite sequence that maps $\{1, \ldots, n+1\}$ into A, has h as an initial subsequence and maps $n+1$ to a.

[19] A partition of $\{h \in H \backslash T : P(h) = i\}$ is a set of nonempty, pairwise disjoint sets $\mathcal{H}_n \subseteq \{h \in H \backslash T : P(h) = i\}$ such that $\bigcup \mathcal{H}_n = \{h \in H \backslash T : P(h) = i\}$.

B Virtual Implementation in Weakly Rationalizable Strategies

In this appendix, we show that for the purpose of virtually wr-implementing social choice functions restricting attention to static mechanisms is without loss of generality. This stands in contrast to results for exact wr-implementation (Sections 3 and 4) and also for virtual implementation in strongly rationalizable strategies (Müller, 2012). For these implementation concepts, dynamic mechanisms can implement strictly more social choice functions than static mechanisms.

In order to characterize virtual wr-implementation we do not need to resort to infinite mechanisms. Hence in this appendix, we restrict attention to *finite* (and thus automatically well-behaved) mechanisms. Trivially, dynamic mechanisms can virtually wr-implement any social choice function that static mechanisms can virtually wr-implement. To show the converse, recall Bergemann and Morris' (2009b) characterization of robust virtual implementation (rv-implementation), an implementation concept that equals virtual wr-implementation in static mechanisms. A key role in their characterization plays strategic indistinguishability. Roughly, two payoff type profiles are strategically indistinguishable if no mechanism provides them with incentives that guarantee that their strategies induce different outcomes. Bergemann and Morris' (2009b) results imply that

- a social choice function is rv-implementable if and only if it is epIC and robustly measurable, that is, if and only if it is epIC and assigns the same outcome to statically strategically indistinguishable payoff type profiles, and that
- two payoff type profiles are statically strategically indistinguishable if and only if they are "inseparable."

Any social choice function that is virtually wr-implementable by dynamic mechanisms must be epIC and must assign the same outcome to wr-strategically indistinguishable payoff type profiles (compare Proposition 5). Here, wr-strategically indistinguishable means strategically indistinguishable by dynamic mechanisms under weak rationalizability. Proposition 4 will imply that two payoff type profiles are wr-strategically indistinguishable exactly if they are inseparable, and therefore exactly if they are statically strategically indistinguishable. Consequently, any virtually wr-implementable social choice function must be robustly measurable. In summary, we obtain that a social choice function is virtually wr-implementable in dynamic mechanisms if and only if it is epIC and robustly measurable, that is, if and only if it is rv-implementable (Corollary 2).

B.1 Inseparability and WR-Strategic Indistinguishability

Bergemann and Morris (2009b) introduced strategic indistinguishability by static mechanisms.

In Müller (2012) we extend their notion to dynamic mechanisms for the solution concept of strong rationalizability. The idea behind *strategic indistinguishability under weak rationalizability (wr-strategic indistinguishability)* is the same as in these papers: two payoff type profiles are strategically indistinguishable if there exists no mechanism in which the observed behavior (= path of play) of the profiles is guaranteed to be different.

Definition 13 *We write $\theta \sim^{\Gamma} \theta'$ and say that the payoff type profiles $\theta \in \Theta$ and $\theta' \in \Theta$ are Γ-wr-strategically indistinguishable if Γ is a mechanism and $\zeta(s) = \zeta(s')$ for some $s \in Q^{\infty}(\theta)$, $s' \in Q^{\infty}(\theta')$. We write $\theta \sim \theta'$ and say that θ and θ' are wr-strategically indistinguishable if $\theta \sim^{\Gamma} \theta'$ for every mechanism Γ.*

Wr-strategic indistinguishability generalizes Bergemann and Morris' original notion of strategic indistinguishability by static mechanisms. In fact, θ and θ' are statically strategically indistinguishable if and only if θ and θ' are Γ-wr-strategically indistinguishable for all static mechanisms Γ.

In order to characterize wr-strategic indistinguishability, we review the idea of inseparability (see Bergemann and Morris, 2009b, 2011). Let $\mathcal{R}$ denote the set of preference relations on Y, that is, the set of complete and transitive binary relations on Y. For $\theta_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i})$ let $R_{\theta_i,\psi_i} \in \mathcal{R}$ denote the preference relation such that for all $y, y' \in Y$

$$yR_{\theta_i,\psi_i}y' \quad \text{iff} \quad E_{\psi_i}u_i(y,\theta) \geq E_{\psi_i}u_i(y',\theta).$$

For $\Psi_{-i} \subseteq \Theta_{-i}$ let

$$\mathcal{R}_i(\theta_i, \Psi_{-i}) = \{R \in \mathcal{R} : R = R_{\theta_i,\psi_i} \text{ for some } \psi_i \in \Delta(\Theta_{-i}) \text{ with } \psi_i(\Psi_{-i}) = 1\}$$

be the set of θ_i's preference relations arising from beliefs supported by Ψ_{-i}. We say that $\Psi_{-i} \subseteq \Theta_{-i}$ *separates* $\Psi_i \subseteq \Theta_i$ if

$$\bigcap_{\theta_i \in \Psi_i} \mathcal{R}_i(\theta_i, \Psi_{-i}) = \emptyset,$$

that is, if not all of the payoff types in Ψ_i can have a preference relation in common whenever agent i believes that $-i$'s payoff types are in Ψ_{-i}. Using this definition, for all $i \in \mathcal{I}$ we let $\Xi_i^0 = 2^{\Theta_i}$ be the set of all subsets of Θ_i and recursively define the set of $(k+1)$-inseparable subsets of Θ_i by

$$\Xi_i^{k+1} = \{\Psi_i \in \Xi_i^k : \exists \Psi_{-i} \in \Xi_{-i}^k, \Psi_{-i} \text{ does not separate } \Psi_i\},$$

$k \in \mathbb{N}$. Then $\Xi_i^{\infty} = \bigcap_{k \in \mathbb{N}} \Xi_i^k$ is the set of all inseparable subsets of Θ_i . Finally, we say that

the payoff type profiles $\theta, \theta' \in \Theta$ are *inseparable* if $\prod_{i \in \mathcal{I}} \{\theta_i, \theta_i'\} \in \Xi^\infty$.

Proposition 4 shows that inseparable payoff type profiles are wr-strategically indistinguishable. Its proof generalizes Bergemann and Morris' (2009b) proof for the static case.

Proposition 4 *If $\theta \in \Theta$ and $\theta' \in \Theta$ are inseparable then $\theta \sim \theta'$.*

Proof. We are going to prove a slightly stronger claim: if $\theta, \theta' \in \Theta$ are inseparable payoff type profiles, then $Q^{\infty,\Gamma}(\theta) \cap Q^{\infty,\Gamma}(\theta') \neq \emptyset$ for any mechanism Γ.

Take any mechanism Γ. We claim that for each $k \in \mathbb{N}$ there exists, for each $i \in \mathcal{I}$ and $\Psi_i \in \Xi_i^\infty$, a strategy $s_i^k(\Psi_i) \in S_i$ such that $(s_i^k(\Psi_i), \theta_i) \in W_i^k$ for all $\theta_i \in \Psi_i$. Obviously this claim holds for $k = 0$. Suppose now it holds for $k' \in \mathbb{N}$. Fix an arbitrary $i \in \mathcal{I}$ and $\Psi_i \in \Xi_i^\infty$. Since for some $\hat{k} \in \mathbb{N}$, $\Xi_i^\infty = \Xi_i^k$ for all $k \geq \hat{k}$, there exists $\Psi_{-i} \in \Xi_{-i}^\infty$ that does not separate Ψ_i. Hence, letting $R \in \bigcap_{\theta_i \in \Psi_i} \mathcal{R}_i(\theta_i, \Psi_{-i})$, for each $\theta_i \in \Psi_i$ there exists $\psi_i^{\theta_i} \in \Delta(\Theta_{-i})$, $\psi_i^{\theta_i}(\Psi_{-i}) = 1$, such that $R_{\theta_i, \psi_i^{\theta_i}} = R$. Let $m_i : \bar{\mathscr{H}}_i \to S_{-i}$ be such that

- $m_i(\{\emptyset\}) = s_{-i}^{k'}(\Psi_{-i})$,
- $m_i(\mathcal{H}) \in S_{-i}(\mathcal{H})$ for all $\mathcal{H} \in \bar{\mathscr{H}}_i$,
- if $\mathcal{H}' \preceq \mathcal{H}$ and $m_i(\mathcal{H}') \in S_{-i}(\mathcal{H})$, then $m_i(\mathcal{H}) = m_i(\mathcal{H}')$.

For each $\theta_i \in \Psi_i$ we let $\mu_i^{\theta_i}$ be the CPS that at information set $\mathcal{H} \in \bar{\mathscr{H}}_i$ prescribes the belief which puts marginal probability one on $m_i(\mathcal{H})$ and has $\psi_i^{\theta_i}$ as marginal distribution on the opponents' payoff types.[20] Pick a $\hat{\theta}_i \in \Psi_i$ and a $s_i \in r_i(\hat{\theta}_i, \mu_i^{\hat{\theta}_i})$. Then

$$\sum_{\theta_{-i} \in \Theta_{-i}} \psi_i^{\hat{\theta}_i}(\theta_{-i}) u_i(C(\zeta(s_i, m_i(\mathcal{H}))), (\hat{\theta}_i, \theta_{-i})) \geq \sum_{\theta_{-i} \in \Theta_{-i}} \psi_i^{\hat{\theta}_i}(\theta_{-i}) u_i(C(\zeta(s_i', m_i(\mathcal{H})), (\hat{\theta}_i, \theta_{-i}))$$

for all $\mathcal{H} \in \mathscr{H}_i(s_i)$ and all $s_i' \in S_i(\mathcal{H})$. By construction, payoff type θ_i with CPS $\mu_i^{\theta_i}$ has preferences R over Y, for all $\theta_i \in \Psi_i$ and all $\mathcal{H} \in \mathscr{H}_i$. Hence if we let $s_i^{k'+1}(\Psi_i)$ equal s_i then $s_i^{k'+1}(\Psi_i) \in r_i(\theta_i, \mu_i^{\theta_i})$ for all $\theta_i \in \Psi_i$. Because $\mu_i^{\theta_i} \in \Pi_i^{k'}$ for all $\theta_i \in \Psi_i$, this implies $(s_i^{k'+1}(\Psi_i), \theta_i) \in W_i^{k'+1}$ for all $\theta_i \in \Psi_i$. This completes the proof by induction, and since there exists $\hat{k} \in \mathbb{N}$ such that $W^k = W^\infty$ for all $k \geq \hat{k}$, we actually proved that there exists, for each $i \in \mathcal{I}$ and $\Psi_i \in \Xi_i^\infty$, a strategy $s_i^\infty(\Psi_i) \in S_i$ such that $(s_i^\infty(\Psi_i), \theta_i) \in W_i^\infty$ for all $\theta_i \in \Psi_i$. $\square$

If two payoff type profiles are inseparable then they are wr-strategically indistinguishable, and hence statically strategically indistinguishable. Bergemann and Morris (2009b, Theorem 1) show that if the agents' preferences satisfy the following no complete indifference condition

[20] Formally, let $\mu_i^{\theta_i} : 2^{\Sigma_{-i}} \times \bar{\mathscr{H}}_i \to [0,1]$ satisfy $\mu_i^{\theta_i}((m_i(\mathcal{H}), \theta_{-i}) | \mathcal{H}) = \psi_i^{\theta_i}(\theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$ and $\mathcal{H} \in \bar{\mathscr{H}}_i$.

then any two statically strategically indistinguishable payoff type profiles are inseparable.[21] This proves the upcoming corollary.

Definition 14 (No Complete Indifference.) *The no complete indifference condition is satisfied if for each $i \in \mathcal{I}$, $\theta_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i})$ there exist $x, x' \in X$ such that*

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

Corollary 1 *If the no complete indifference condition is satisfied then $\theta \sim \theta'$ if and only if θ and θ' are inseparable (and therefore, if and only if θ and θ' are statically strategically indistinguishable).*

B.2 Necessary and Sufficient Conditions for Virtual WR-Implementation

A social choice function is virtually wr-implementable if it can be approximately wr-implemented in the following sense, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^{\#X}$.

Definition 15 *Social choice function f is ε-wr-implementable if there is a finite mechanism Γ such that $\|C(\zeta(s)) - f(\theta)\| \leq \varepsilon$ for all $(s, \theta) \in W^\infty$. Social choice function f is virtually implementable in weakly rationalizable strategies (virtually wr-implementable) if it is ε-wr-implementable for every $\varepsilon > 0$.*

Ex-post incentive compatibility (see Definition 9) is necessary for robust and robust virtual implementation in static mechanisms (see Bergemann and Morris, 2005, 2009b, 2011). The same is true for exact wr-implementation (Proposition 2) and, as Proposition 5 clarifies, for virtual wr-implementation in dynamic mechanisms. Proposition 5 furthermore establishes robust measurability as a second necessary condition for virtual wr-implementation. Recall that Bergemann and Morris (2009b) call a social choice function f *robustly measurable* if $f(\theta) = f(\theta')$ whenever $\theta \in \Theta$ and $\theta' \in \Theta$ are inseparable.

Proposition 5 *Suppose the no complete indifference condition is satisfied. If social choice function f is virtually wr-implementable, then f is epIC and robustly measurable.*

Proof. Suppose f is virtually wr-implementable. We first show that f is robustly measurable. Take $\varepsilon > 0$, then there is a mechanism Γ that ε-wr-implements f. Suppose θ and θ' are inseparable, then by Corollary 1, $\theta \sim \theta'$. Hence there are $s \in Q^\infty(\theta)$ and $s' \in Q^\infty(\theta')$ such that $\zeta(s) = \zeta(s')$. By ε-wr-implementation, $\|C(\zeta(s)) - f(\theta)\| \leq \varepsilon$ and $\|C(\zeta(s')) - f(\theta')\| \leq \varepsilon$ and thus $\|f(\theta) - f(\theta')\| \leq 2\varepsilon$. Since this is true for all $\varepsilon > 0$, $f(\theta) = f(\theta')$.

[21] Bergemann and Morris (2009b) also assume that all agents' payoff type spaces Θ_i have the same cardinality, but this assumption is for convenience only.

Müller (2012, Proposition 1) implies that f is epIC.[22] □

Bergemann and Morris (2009b) show that if the following economic property (which implies the no total indifference condition) is satisfied then a social choice function is rv-implementable if and only if it is epIC and robustly measurable. Let $\bar{y}$ denote the uniform lottery placing probability $\frac{1}{\#X}$ on each $x \in X$.

Definition 16 (Economic Property) *The economic property is satisfied if there exists a profile of lotteries $(z_i)_{i\in\mathcal{I}} \in Y^I$ such that for each $i \in \mathcal{I}$ and $\theta \in \Theta$ both $u_i(z_i, \theta) > u_i(\bar{y}, \theta)$ and $u_j(\bar{y}, \theta) \geq u_j(z_i, \theta)$, $j \neq i$.*

If f is rv-implementable, then it trivially is virtually wr-implementable. Therefore we obtain the following characterization of virtual wr-implementation.

Corollary 2 *Suppose the economic property is satisfied. Then social choice function f is virtually wr-implementable if and only if f is epIC and robustly measurable (and therefore, if and only it is rv-implementable).*

References

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., and M. Siniscalchi (1999): "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games," *Journal of Economic Theory*, 88, 188–230.

——— (2002): "Strong Belief and Forward Induction Reasoning," *Journal of Economic Theory*, 106, 356–391.

——— (2003): "Rationalization and Incomplete Information," *Advances in Theoretical Economics*, 3(1), Article 3.

Ben-Porath, E. (1997): "Rationality, Nash Equilibrium and Backwards Induction in Perfect-Information Games," *Review of Economic Studies*, 64, 23–46.

Bergemann, D., and S. Morris (2005): "Robust Mechanism Design," *Econometrica*, 73(6), 1771–1813.

[22] In Müller (2012, Proposition 1) we prove that epIC is necessary for virtual implementation in strongly rationalizable strategies. If f is virtually wr-implementable, then f is virtually implementable in strongly rationalizable strategies, and thus epIC.

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

——— (2009b): "Robust virtual implementation," *Theoretical Economics*, 4(1), 45–88.

——— (2011): "Robust implementation in general mechanisms," *Games and Economic Behavior*, 71(2), 261–281.

Bikhchandani, S. (2006): "Ex post implementation in environments with private goods," *Theoretical Economics*, 1, 369–393.

Brandenburger, A., and E. Dekel (1987): "Rationalizability and Correlated Equilibria," *Econometrica*, 55(6), 1391–1402.

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

Dasgupta, P., and E. S. Maskin (2000): "Efficient Auctions," *Quarterly Journal of Economics*, 115(2), 341–388.

Harsanyi, J. C. (1967-68): "Games with Incomplete Information Played by "Bayesian" Players, Parts I-III," *Management Science*, 14, 159–182, 320–334, 486–502.

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

Jehiel, P., M. Meyer-Ter-Vehn, B. Moldovanu, and W. R. Zame (2006): "The Limits of Ex Post Implementation," *Econometrica*, 74(3), 585–610.

Kuhn, H. W. (1953): "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games, Vol. II*, ed. by H. W. Kuhn, and A. W. Tucker, vol. 28 of *Annals of Mathematics Studies*, pp. 193–216. Princeton University Press.

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

Müller, C. (2010): "Robust Implementation in Dynamic Mechanisms," Ph.D. thesis, University of Minnesota.

Müller, C. (2012): "Robust Virtual Implementation under Common Strong Belief in Rationality," Working Paper.

Neeman, Z. (2004): "The Relevance of Private Information in Mechanism Design," *Journal of Economic Theory*, 117, 55–77.

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

Penta, A. (2009): "Robust Dynamic Mechanism Design," Working Paper.

Perea, A. (2011): "Belief in the Opponents' Future Rationality," Working Paper.

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(3), 385–391.

Tan, T. C.-C., and S. R. d. C. Werlang (1988): "The Bayesian Foundations of Solution Concepts of Games," *Journal of Economic Theory*, 45, 370–391.

Weinstein, J., and M. Yildiz (2007): "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements," *Econometrica*, 75(2), 365–400.

Wilson, R. (1987): "Game-theoretic analyses of trading processes," in *Advances in Economic Theory: Fifth World Congress*, ed. by T. F. Bewley, vol. 12 of *Econometric Society Monographs*, chap. 2, pp. 33–70. Cambridge University Press.