# On Necessary Conditions for Implementation of Functions, without Rational Expectations

Giacomo Rubbini*

Preliminary and Incomplete

Please do not cite or circulate

March 25, 2022

Link to the latest version

**Abstract**

The Bayesian implementation literature has identified in Bayesian Incentive Compatibility (BIC) and Bayesian Monotonicity (BM) two key conditions that a social choice function has to satisfy to be fully implemented by a social planner. I characterize the class of solution concepts such that BIC is necessary for full implementation of functions, and I find we can not expect significantly more permissive results by dropping the rational expectations assumption and moving to non-equilibrium models. Preliminary results suggest the same may be true for a BM-like condition as well.

**Keywords:** Mechanism Design, Bounded Rationality, Rational Expectations

**JEL:** C72, D82, D90

*Department of Economics, Brown University, giacomo_rubbini@brown.edu

# 1 Introduction

Can a social planner implement a given social goal by designing rules of interaction between agents, when these agents hold private information they can exploit to their advantage? As the answer to such a question clearly depends on how such an interaction pans out, the mechanism design and implementation literature has extensively explored this problem using a variety of game-theoretic solution concepts. One of the leading ones in this sense is Bayesian Nash Equilibrium (BNE): Jackson (1991), for example, provides a characterization of the set of social choice rules implementable in BNE via three key properties (Bayesian Incentive Compatibility, Bayesian Monotonicity and closure), and Kunimoto (2019) extends this result to mixed BNE[1].

All equilibrium solution concepts, however, rely on the crucial assumption agents are able to correctly anticipate their opponents' strategies, i.e. that agents have rational expectations. Insights from the experimental and behavioral literature have highlighted this assumption is likely not to hold in many settings: for example, when agents face a mechanism for the first time. It is thus natural to wonder whether some of the alternative models that better describe agents' behavior allow for implementation of a class of social choice functions (SCFs) larger than the one that are implementable in BNE[2]. However, recent results about non-equilibrium solution concepts such as level-k reasoning and interim correlated rationalizability suggest the answer to this question may be mostly negative (de Clippel et al., 2019; Kunimoto et al., 2020)[3].

It is still unclear whether considering implementation in solution concepts different from the ones mentioned above may yield different insights from rational expectations models and, if so, which solution concepts we should consider. I pro-

[1]See Maskin and Sjöström (2002) or Serrano (2004) for an overview of the literature.

[2]The rational expectations assumption is used in behavioral implementation models as well. For example, de Clippel (2014) extends the classical results from Maskin (1999) by relaxing preference maximization, but it still assumes agents correctly anticipate the set of opportunities they will be choosing from.

[3]Kneeland (2022) obtains more permissive results in the context of level-k implementation of social choice *sets*.

vide an answer to this question by turning on its head implementation theory's standard approach of fixing a solution concept and then proceeding to characterize the set of restrictions SCFs have to satisfy in order to be implementable. Rather than checking necessary conditions for each solution concept separately, I characterize the class of *all* solutions concepts such that Bayesian Incentive Compatibility (BIC) is necessary for implementation.

My results indeed confirm that BIC is still necessary for full implementation of functions without rational expectations, as long as the solution concept used satisfies a mild condition I call Weak Best Response Consistency (WBRC). WBRC requires that, for each agent and pair of types, there exists at least one solution of the implementing mechanism such that the agent prefers to play the strategy associated to her type rather than mimicking a different one. A sufficient condition for WBRC is that each agent expects her opponents to be best responding to their (not necessarily rational) expectations. Crucially, necessity of BIC does not rely on the assumption of common knowledge of rationality. In Section 4.2 I show both the models of de Clippel et al. (2019) and Kunimoto et al. (2020) satisfy this condition for any given mechanism, but the same is not true for Eyster and Rabin (2005)'s cursed equilibrium solution concept for the case of common values (Appendix A).

The results about BIC suggest that we can generally not expect to expand dramatically the set of implementable SCFs by moving to non-equilibrium solution concepts. I am moreover able to prove that if a BIC social choice function is implementable, then WBRC holds. By providing a full characterization of the set of solution concepts that allow implementation of non-BIC SCFs, I provide some guidance as to which solutions concepts may successfully escape necessity of BIC.

This paper also relates to the literature studying how sensible results in mechanism design are with respect to changes in the details of the planner's model of the players, and in particular to the literature about continuous (Oury and Tercieux, 2012; de Clippel et al., 2021) and robust (Bergemann and Morris, 2005, 2011) implementation. These approaches can be interpreted as focusing on a planner having an imperfect knowledge (or no knowledge at all) about the agents' payoffs and beliefs. The interpretation of my approach may instead be better

understood as considering a planner who has an accurate model of payoffs and beliefs, but who is not sure how these maps into what players expect from their opponents. In this sense, my framework can be thought of as studying how sensible some restrictions on the set of implementable SCFs (such as BIC) are with respect to alterations of the model of expectation formation rather than of payoffs and beliefs.

# 2 Model

The goal of the social planner is to select an alternative from a set $A$, conditional on some information privately held from the agents in set $I$. The incomplete information problem by assuming there exists a set of types $T_i$ available to each agent $i \in I$, and that each agent knows her type but not the type of other players. Let moreover $T = \times_{i \in I} T_i$ be the set of all possible type profiles.

Agents' (interim) beliefs about the types of their opponents are denoted as $p_i : T_i \to \Delta(T_{-i})$: that is, when an agent is of type $t_i$, she believes other players are of types $t_{-i}$ with probability $p(t_{-i}|t_i)$[4]. Let me assume that for all $t \in T$ there exists at least one $i \in I$ such that $p(t_{-i}|t_i) > 0$ [5]. I assume preferences over lotteries have expected utility form, with Bernoulli utility $u_i : A \times T \to \mathbb{R}$. Abusing slightly of notation, let $u_i(a, t)$ for $a \in \Delta(A)$ denote the utility agent $i$ derives from lottery $a$ when the type profile is $t$.

The social planner seeks to implement a social choice function $f : T \to \Delta(A)$, and she does so by designing a mechanism $\gamma = (\mu, S)$ where $S = \times_{i \in I} S_i$ is an action space and $\mu : S \to \Delta(A)$ is an outcome function. Let $\Gamma$ denote the set of all possible mechanisms the planner can design. Once the planner has committed to a mechanism, agents choose a strategy $\sigma_i : t_i \to \Delta(S_i)$. We will denote the set of all such functions as $\Sigma_i$ and a profile of strategies $\{\sigma_i\}_{i \in I}$ as $\sigma \in \Sigma$. For

[4]For example, we can take $p(t_{-i}|t_i)$ to be the Bayesian posterior stemming from a common prior distribution $q : T \to (0, 1)$.

[5]This assumption is not necessary to get to the results I will present, but it will make notation more convenient. In particular, it will not be necessary to state my results in terms of equivalent SCFs.

the rest of the paper, let me slightly abuse of the notation above by considering $\mu(\sigma(t))$ to denote the lottery over $A$ induced by $\sigma(t)$ under outcome function $\mu$.

A key feature of rational expectations is that agents' expectations turn out to be correct in equilibrium: for example, if $\sigma$ is a BNE, player $i$ expects her opponents to play $\sigma_{-i}$. In order to relax the rational expectations assumption, we will instead consider a more general model of agents' expectations. For a given mechanism $\gamma$, let $e_{t_i} \in \Sigma_{-i}$ represent the expectations of type $t_i$ of agent $i$ about her opponent. Notice that, by this definition, $e_{t_i}$ belongs to the strategy space for players $j \neq i$: that is, $e_{t_i} \in \Sigma_{-i}$[6] As $e_{t_i}$ is a strategy profile for players $j \neq i$, we will sometimes evaluate it at $t_{-i}$: thus, $e_{t_i}(t_{-i}) \in \Delta(S_{-i})$. To make notation more compact, let me define a mapping $e_i : T_i \to \Sigma_{-i}$ that assigns $e_{t_i}$ to each type $t_i \in T_i$ and let me denote as $e$ any profile $(e_i)_{i \in I}$. For $j \neq i$, I will furthermore denote as $e_i^j$ and $e_{t_i}^j$ the $j$-th entry of (respectively) $e_i$ or $e_{t_i}$. In other words, $e_{t_i}^j$ corresponds to the strategy $t_i$ expects $j$ to play: this entails $e_{t_i}^j \in \Sigma_j$. The set of all possible expectations over mechanism $\gamma$ will be denoted as $\mathcal{E}(\gamma) = (\Sigma_{-i})_{i \in I}$.

Let then a *model of expectations* $E$ be any correspondence mapping each mechanism $\gamma$ into a subset $E(\gamma)$ of $\mathcal{E}(\gamma)$. We will interpret $E(\gamma)$ as the expectations the model allows agents to hold: for example, interim correlated rationalizability implicitly rules out the possibility agents expect one of their opponent to play a dominated strategy (the reader can refer to Section 4.2 for some examples of models of expectations). As in de Clippel et al. (2019) and Kunimoto et al. (2020), we can interpret $E(\gamma)$ as the set of expectations the planner believes will I assume the planner sees each of this expectation profiles as happening with non-zero probability. This will be reflected in the implementation concept used below, which requires the outcome prescribed by $f$ to prevail regardless of the expectation profile considered.

We will moreover denote as $B_{t_i'}(e_{t_i})$ the set of best replies for type $t_i'$ to the

[6]This formulation implicitly assumes expectations are deterministic: however, given we assume agents' preferences over lotteries admit an expected utility representation, this assumption does not cause further loss of generality.

expectations of type $t_i$[7]. That is, if $\sigma_i(t_i) \in B_{t_i'}(e_{t_i})$, then for all $s_i' \in S_i$:

$$\int_{t_{-i}} u_i(\mu(\sigma_i(t_i), e_i(t_{-i}), t')\, dp_i(t_{-i}|t_i') \geq \int_{t_{-i}} u_i(\mu(s_i', e_i(t_{-i}), t')\, dp_i(t_{-i}|t_i')$$

To keep notation compact, let $\sigma_i \in B_i(e_i)$ mean be the set of $\sigma \in \Sigma$ such $\sigma_i(t_i) \in B_{t_i}(e_{t_i})$ for all $t_i \in T_i$, $\sigma \in B(e)$ the set of $\sigma$ such that $\sigma_i \in B_i(e_i)$ for $i \in I$ and $B_t(e_t)$ the set of $\sigma$ such that $\sigma_i(t_i) \in B_{t_i}(e_{t_i})$ for all $i \in I$.

Let me now define the implementation concept I will use in the remainder of the paper. Let $\sigma$ be a *solution* to a mechanism $\gamma$ whenever $\sigma \in B(e)$ for $e \in E(\gamma)$[8]. We will say a mechanism $\gamma$ fully implements $f$ whenever $\gamma$ has at least one solution and every such solutions yields the outcome prescribed by $f$. Formally:

**Definition 1.** *We say a SCF $f$ is fully implementable given $E$ whenever there exists $\gamma$ such that:*

- $B(e) \neq \emptyset$ *for all* $e \in E(\gamma)$
- $f = \mu \circ \sigma$ *for all profiles* $\sigma$ *such that there exists* $e \in E$ *with* $\sigma \in B(e)$

In the remainder of the paper, let me refer to "full implementation" simply as "implementation" unless otherwise specified.

Before moving on to next section, let me add a few additional technical assumptions. To make sure expected utility is well defined over the spaces discussed in the paper, let $A$, $T_i$ and $S_i$ be separable metrisable spaces endowed with the Borel sigma algebra, let product sets be endowed with the product topology, the Bernoulli utility functions be bounded and continuous, and SCFs, mechanisms, and strategies be measurable functions.

[7]The set of best responses should depend on the specific mechanism used as well, but I omit that to make notation lighter.

[8]This definition of a solution does not include BNE as a special case, as it requires *all* best replies to a profile of expectations to be solutions of the mechanism (see Section 4.2.3 for a comparison of the two concepts). The results about BIC extend with very similar intuition to a more general model that is able to do so Appendix A.

# 3 A bilateral trading example

In order to make the intuition behind necessity of BIC, I will borrow the example of bilateral trade between level-k parties from Crawford (2021) and its discussion in de Clippel et al. (2019).

Before moving to the example itself, let me quickly summarize how level-k models of behavior work. Level-0 players of type $t_i$ are considered to be naïve and to (non-strategically) play some anchor $\alpha_i(t_i)$, which is taken as exogenous to the model. Level-1 agents, instead, are assumed to believe their opponents are level-0 and to be best responding to these opponents' anchors. We will say any such best response is a *level-1 consistent* strategy, denoted as $\sigma^1$. For every level $k_i > 1$, agents of level $k$ believe their opponents to be playing a *level-(k-1) consistent* strategy $\sigma^{k_i-1}$, and best respond accordingly. We say a profile $\sigma$ is a solution to a game $\gamma$ whenever there exists a combination of levels $\{k_i\}_{i\in I}$ such that $k_i > 0$ for all $i \in I$ and $\sigma_i$ is level-$k_i$ consistent for all $i \in I$.

Suppose now two risk-neutral parties trade an indivisible object which has value $c$ for the seller and $v$ for the buyer, both distributed uniformly between 0 and 1. They trade using as a protocol a $\frac{1}{2}$-double auction: the seller and the buyer respectively submit an ask $a$ and a bid $b$ for the object, and trade happens if and only the $b \geq a$, at price $p = 0.5(a + b)$. Utility from not trading is 0 for both parties, while the utility from trading is $u_s = p - c$ and $u_b = b - p$ for the seller and the buyer respectively.

In the discussion in Crawford (2021), it is assumed the agents' anchor is uniformly distributed over $[0, 1]$ and that both agents are of level $k = 1$. Then there exists a SCF $f$ that is implementable but not BIC: the unique level-1 consistent strategies are to bid $\frac{2}{3}v$ for the buyer and to ask $\frac{1}{3}c + \frac{1}{3}$ for the seller, and the associated SCF stipulates trade happens if and only if $2v \geq 1 + c$ at a price of $\frac{1}{6}(2v + c + 1)$. As remarked by de Clippel et al. (2019), a buyer of value $v = 0.5$ would then have an incentive to imitate a buyer of type $v = 0.75$ to gain a payoff of $0.75 - \frac{1}{6}(2 + c)$ rather than 0, violating Bayesian Incentive Compatibility.

However, the same function is not implementable if the two agents could both be of level $k = 2$. As a matter of fact, de Clippel et al. (2019) highlight that

playing $\frac{2}{3}v + \frac{1}{9}$ for $v \geq \frac{1}{3}$ and $v$ otherwise is a best response for the buyer to the level-1 strategy of the seller. Similarly, playing $\frac{2}{3}c + \frac{2}{9}$ for $c \geq \frac{1}{3}$ and $c$ otherwise is a best reply for the seller to a level-1 buyer. The strategies form then a solution to the mechanism considered: the mechanism fails to implement $f$, however, as the two solutions lead to different outcomes. As a matter of fact, it is straightforward to check that trade would now take place for any values of $v, c$ such that $\frac{1}{3} > v \geq c$: this is not the case for level-1 players, who never trade for $v < \frac{1}{2}$.

This discrepancy follows in this case from the fact that we need *all* solutions of the mechanism to yield the same outcome for each type profile $t \in T$ in order for the mechanism to implement a SCF $f$. The argument, however, generalizes to any arbitrary mechanism $\gamma = (\mu, S)$: suppose $\mu$ has a solution $\sigma^1$ (so that $\sigma_i^1$ is a best reply to $\alpha_{-i}$ for all agents) and that such a solution induces a non-incentive compatible SCF. Then, $(\sigma_i^2, \sigma_{-i}^1)$ is a solution of the mechanism as well for any best response $\sigma_i^2$ to $\sigma_{-i}^1$, as player $i$ is best responding to level-1 consistent strategies while all other agents are best responding to their anchors. Moreover, it cannot be that $\mu(\sigma^1) = \mu(\sigma^2)$: as $\sigma^1$ induces non-BIC $f$, there would then exist $i \in I$, $t_i, t_i' \in T_i$:

$$\begin{aligned}
&\int_{t_{-i}} u_i(\mu(\sigma_i^2(t_i'), \sigma_{-i}^1(t_{-i})), t)\, dp(t_{-i}|t_i) = \\
&\int_{t_{-i}} u_i(\mu(\sigma_i^1(t_i'), \sigma_{-i}^1(t_{-i})), t)\, dp(t_{-i}|t_i) > \\
&\int_{t_{-i}} u_i(\mu(\sigma_i^1(t_i), \sigma_{-i}^1(t_{-i})), t)\, dp(t_{-i}|t_i) = \\
&\int_{t_{-i}} u_i(\mu(\sigma_i^2(t_i), \sigma_{-i}^1(t_{-i})), t)\, dp(t_{-i}|t_i)
\end{aligned}$$

But this implies $\sigma^2$ is not a best reply to $\sigma^1$ for at least one type $t_i$ of player $i$. It must then be that $\mu(\sigma^1) \neq \mu(\sigma^2)$: this violates uniqueness, making it impossible for the mechanism to implement any SCF[9].

This insight seems to rely on the properties of level-k models, that are often solved recursively starting from the anchor. Quite surprisingly, this is not the case:

[9]It would still be possible for the mechanism to implement a social choice *set*: as a matter of fact, Kneeland (2022) proves BIC is no longer necessary for (partial) level-k implementation in this case.

in the next sections, we will prove the same holds true for a much larger class of solutions concepts. In particular, any model in which players are best responding to one another and believe others to be best responding to their expectations as well makes BIC necessary for implementation. This class of solutions concepts is larger than the class of equilibrium ones, as each player and type may be best responding to expectations that are different from the ones of their opponents and of other types.

# 4 Necessity of BIC

I will now characterize the class of solution concepts such that a mild strengthening of BIC (SIRBIC) is necessary for implementation of functions[10].

**Definition 2** ((Strict-if-Responsive) Bayesian Incentive Compatibility)**.** *We say a SCF is BIC whenever for all $i \in I$ and $t_i, t_i' \in T_i$:*

$$\int_{t_{-i}} u_i(f(t), t)\, dp_i(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(f(t_i', t_{-i}), t)\, dp_i(t_{-i}|t_i)$$

*If moreover the inequality is strict for all pairs $t_i, t_i'$ such that $f(t_i', t_{-i}) \neq f(t)$ for some $t_{-i} \in T_{-i}$, we say $f$ is SIRBIC.*

The class of models that lead to necessity of SIRBIC can be characterized by imposing the following restriction on the solution of the implementing mechanism.

**Definition 3** (Weak Best Response Consistency (WBRC))**.** *We say a model of expectations $E$ satisfies WBRC for mechanism $\gamma$ whenever for all $i \in I$, $t_i, t_i' \in T_i$ there exist $e \in E(\gamma)$ and $\sigma \in B(e)$ such that:*

$$\int_{t_{-i}} u_i(\mu(\sigma(t), t)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(\mu(\sigma(t_i', t_{-i}), t)\, dp(t_{-i}|t_i)$$

To get a first intuition about why we need such a property for the result of Theorem 1, consider first Bayesian implementation. For each equilibrium $\sigma$,

[10]The argument in the proof of Theorem 1 actually proves that any $f$ that is implementable and BIC is also SIRBIC. See the more general model of Appendix A for further discussion.

type $t_i$ is able to induce either lottery $\mu(\sigma_i(t_i), \int_{t_{-i}} \sigma_{-i}(t_{-i})\, dp_i(t_{-i}|t_i))$ or lottery $\mu(\sigma_i(t_i'), \int_{t_{-i}} \sigma_{-i}(t_{-i})\, dp_i(t_{-i}|t_i))$ by changing her action from $\sigma(t_i)$ to $\sigma(t_i')$. As $\sigma$ is a solution of the mechanism, this is equivalent to a comparison between lotteries $f(t_i, \int_{t_{-i}} t_{-i}\, dp_i(t_{-i}|t_i))$ and $f(t_i', \int_{t_{-i}} t_{-i}\, dp_i(t_{-i}|t_i))$, leading us to necessity of BIC. The same intuition can be preserved for a broader class of solution concepts, as WBRC ensures that type $t_i$ faces the same comparison above whenever $E$ is WBRC for implementing mechanism $\gamma$.

Before moving to the main result of the section, it is worth noticing that the solution $\sigma$ inducing this comparison needs not to be the same for all players $i$. As Kneeland (2022) highlights in her Remark 2, this is due to the fact we do not require agents' expectations to be consistent anymore. The solution needs not to be the same for any type pair $t_i, t_i'$ of player $i$ either: this relaxation follows instead from the fact we allow expectations to be type-dependent[11].

**Theorem 1.** *If, whenever $f$ is implementable given $E$ via mechanism $\gamma$:*

- *$f$ is SIRBIC, then $E$ is WBRC for $\gamma$*
- *$f$ is not SIRBIC, then $E$ is not WBRC for $\gamma$*

In other words, Theorem 1 provides a characterization of all solution concepts such that SIRBIC is necessary for implementation.

Moreover, it follows as an easy corollary to Theorem 1 that $f$ is SIRBIC whenever it is implementable and $E$ is WBRC for all $\gamma \in \Gamma$: this is the case, for example, for the level-k reasoning model (de Clippel et al., 2019; Kneeland, 2022) and for Interim Correlated Rationalizability (Kunimoto et al., 2020).

## 4.1 Sufficient conditions for WBRC

As WBRC may prove hard to check in practice, I will discuss in this section two properties that actually imply WBRC.

[11] Both level-k reasoning de Clippel et al. (2019); Kneeland (2022) and BNE assume agents' expectations are the same for all types of player $i$, while interim correlated rationalizability Kunimoto et al. (2020) allows for type-dependent expectations.

First, $E$ satisfies WBRC for mechanism $\gamma$ whenever one of its solutions is a BNE as well. As a matter of fact, if $\sigma$ is a BNE, by definition for all $i \in I$, $t_i, t_i' \in T_i$:

$$\int_{t_{-i}} u_i(\mu(\sigma_i(t_i), \sigma_{-i}(t_i)), t) \geq \int_{t_{-i}} u_i(\mu(\sigma_i(t_i'), \sigma_{-i}(t_i)), t)$$

And if $\sigma$ is a solution to $\gamma$ it is immediate to derive WBRC holds. We do not even need $\sigma$ to be a BNE: $\sigma_i(t_i)$ needs not to be a best reply for type $t_i$, just a better one than mimicking type $t_i'$ by playing $\sigma_i(t_i')$ [12]. This connection is actually very intuitive, as in the classical model BIC serves to prevent type $t_i$ from impersonating type $t_i'$ in a direct mechanism. It moreover makes it immediate to check if WBRC holds by showing there exists a BNE such that $\sigma \in B(e)$ for some $e \in E(\gamma)$.

Another useful sufficient condition for BRC makes use of the model of expectations $E$ in a more explicit way. In order to lay down its definition, let $E^*(\gamma)$ be the set of $e \in E(\gamma)$ such that $\sigma \in B(e)$ implies that for all $i \in I$ and $t_i \in T_i$ there exists $e' \in E(\gamma)$ such that $(\sigma_i, e_{t_i}) \in B(e')$. That is, for all $\tilde{t} \in T$, $\sigma_i(\tilde{t}_i)$ is a best reply to $e'_{\tilde{t}_i}$ for player $i$ and $e^j_{t_i}(\tilde{t}_j)$ is a best reply to $e'_{\tilde{t}_j}$ for all $j \neq i$. We can interpret this set as the set of expectation profiles that are consistent with a belief in a minimal level of rationality: each type $t_i$ of player $i$ expects her opponent to best respond (and best responds herself) to some possible expectation profile $e'$. Notice that, as in WBRC, we do not require profile $e'$ to be the same for all types and players.

**Definition 4** (Best Response Consistency (BRC))**.** *We say $E$ satisfies BRC for mechanism $\gamma$ whenever $E^*(\gamma) \neq \emptyset$.*

It is immediate to see BRC implies WBRC as, for each $t_i \in T_i$, $\sigma \in B(e)$ implies $(\sigma_i, e_{t_i}) \in B(\tilde{e})$ for $\tilde{e} \in E(\gamma)$.

Therefore, a model of expectations $E$ is BRC for $\gamma$ whenever there exists at least one expectation $e \in E(\gamma)$ such that each type $t_i$ of each player $i$ can justify the profile $(\sigma_i, e_{t_i})$ they expect to prevail as a best reply to some expectation $e'$

[12] Thus the core of my results goes through even when agents do not best respond to their expectations: for example, in the spirit of Simon (1955)'s satisficing.

which is consistent with model $E$: we are just requiring agents to believe all other agents are best responding to an expectation which is consistent with the model.

## 4.2 Examples

Let me now show these conditions are rather common in two examples taken from the literature (namely de Clippel et al. (2019) and Kunimoto et al. (2020) both show SIRBIC is necessary for implementation of SCFs) and in two examples that, to the best of my knowledge, have not been considered yet.

### 4.2.1 Level k model

Let $\alpha : \Gamma \to \Sigma$ be any correspondence assigning a profile of anchors to each mechanism $\gamma \in \Gamma$. Then we can characterize the model of expectations for level-k reasoning as follows:

$$E^K(\alpha, \gamma) = \{e \in \mathcal{E} : e_i \in \{\alpha_{-i}(\gamma)\} \cup \{\cup_{1 \leq k_i \leq K} S_{-i}^{k_i - 1}(\gamma|\alpha)\}, i \in I\}$$

That is, the set of all $e \in \mathcal{E}$ such that each player $i$ expects the remaining players to play the anchor ($e_i \in \alpha_{-i}(\gamma)$) or to best-respond as players of some level $k_i - 1$ ($e_i \in \cup_{1 \leq k_i \leq K} S_{-i}^{k_i - 1}(\mu|\alpha)$). Notice in this case expectations are type-independent, simplifying a bit our analysis.

As for non-emptiness of $E^{K*}$, it is enough to notice:

$$E^{K*} = \{e \in \mathcal{E} : e_i \in S_{-i}^{k_i - 1}(\gamma|\alpha), 2 \leq k_i \leq K, \forall i \in I\}$$

That is, we need to exclude cases in which a player expects her opponents to just play their anchor. To be clearer on why these expectation profiles belong to $E^*(\gamma)$, consider that $\sigma \in B(e)$ for $e \in E^*(\gamma)$ implies $\sigma$ is a profile of best replies to a profile $\{S_{-i}^{k_i - 1}(\gamma|\alpha)\}_{i \in I}$ with $k_i \geq 2$. If $k_i = 2$, this in turn entails that, for all $i \in I$, $\sigma_i$ is a best reply to $S_{-i}^1(\gamma|\alpha)$ and $e_i^j$ is a best reply to $\alpha_{-j}$: therefore, $(\sigma_i, e_i) \in B(e')$ for $e_i' = S_{-i}^{k_i - 1}(\gamma|\alpha)$ and $e_j = \alpha_{-j}$ for $j \neq i$. A similar argument follows for the case $k_i > 2$, with the difference that now $e_i^j$ is a best reply to $S_{-j}^{k_i - 2}(\gamma|\alpha)$: we can then construct $e'$ by setting $e_i' = S_{-i}^{k_i - 1}(\gamma|\alpha)$ and $e_j = S_{-j}^{k_i - 2}(\gamma|\alpha)$.

We can make sure $E^{K*} \neq \emptyset$ by allowing the possibility all players are at least of level 2: in de Clippel et al. (2019) model, this corresponds to the case in which the upper bound $K$ on the level of the players is such that $K \geq 2$. This seems to account for the differences between DSS and Crawford (2021): the latter proves it is possible to implement non-incentive compatible rules, but that is due to the fact that, by supposing there are no players of level 2, they implicitly assume $E^{K*} = \emptyset$.

### 4.2.2 Rationalizability

Kunimoto et al. (2020) study implementation using Interim Correlated Rationalizability Dekel et al. (2007) as a solution concept, finding that SIRBIC is a necessary condition for implementation of SCFs.

Let $R = (R_i)_{i \in I}$ be a correspondence profile such that for all $i \in I$ we have $R_i : T_i \to 2^{S_i}$. Consider the operator $b = (b_i)_{i \in I}$ iteratively eliminating strategies that are never a best response:

$$b_i(R)[t_i] \equiv \left\{ m_i : \begin{array}{l} \exists \lambda_i \in \Delta(T_{-i} \times S_{-i}) \text{ such that:} \\ (1)\ \lambda_i(t_{-i}, s_{-i}) > 0 \Rightarrow s_{-i} \in R_{-i}(t_{-i}); \\ (2)\ \text{marg}_{T_{-i}} \lambda_i = p_i(t_{-i}|t_i); \\ (3)\ s_i \in \arg\max_{s_i'} \int_{(t_{-i}, s_{-i})} u_i(\mu(s_i', s_{-i}), (t_i, t_{-i}))\, d\lambda_i(t_{-i}, s_{-i}) \end{array} \right\}$$

As argued in the paper, by Tarski's theorem, there exists a largest fixed point of $b$ which is denoted as $R^{\gamma(T)}$. The authors then require, for $f$ to be implementable, that there exists a mechanism such that (1) the desired outcome obtains for all rationalizable strategy profiles and (2) that each type $t_i$ has at least one rationalizable action.

We can then prove this is equivalent to assuming the following theory of expectations:

$$E^I = \{e \in \mathcal{E} : \text{supp}(e_{t_i}(t_{-i})) \subseteq R_{-i}^{\gamma(T)}(t_{-i})\}$$

Notice in this case $E^{I^*} = E^I$, as each strategy in $R_{-i}^{\gamma(T)}(t_{-i})$ is the best response to a rationalizable profile of strategies for the other players. To be clearer, consider any $e \in E^I$: we will show then that $E(\gamma) \subseteq E^*(\gamma)$ as we know by the definition

of $E^*$ that $E^*(\gamma) \subseteq E(\gamma)$. Suppose $\sigma \in B(e)$ and consider, for arbitrary $i \in I$ and $t_i \in T_i$, $(\sigma_i, e_{t_i})$. We then construct $e' \in E(\gamma)$ such that $(\sigma_i, e_{t_i}) \in B(e')$ as follows. We set $e'_i = e_i$. For $j \neq i$ and all $s_j \in S_j$, let $e_{t_j}^{s_j} \in B_{t_j}^{-1}(s_j)$ and set:

$$e'_{t_j} = \int_{s_j} e_{t_j}^{s_j} \, de_{t_i}(t_j)[s_j]$$

We can now show $e' \in E(\gamma)$. If $j \neq i$, $s_{-j} \in \text{supp}(e'_{t_j}(t_{-j}))$ implies that there exists $s_j$ to which $e_{t_i}(t_j)$ assigns non-zero probability and such that $s_{-j} \in \text{supp}(e_{t_j}^{s_j}(t_{-j}))$, where $\text{supp}(e_{t_j}^{s_j}(t_{-j})) \subseteq R^{\gamma(T)}(t_{-j})$: then $\text{supp}(e'_{t_j}(t_{-j})) \subseteq R^{\gamma(T)}(t_{-j})$. As for $i$, $\text{supp}(e'_{t_i}(t_{-i})) = \text{supp}(e_{t_i}(t_{-i})) \subseteq R^{\gamma(T)}(t_{-i})$ by construction. It follows $e' \in E(\gamma)$ and $e \in E^*(\gamma)$, concluding the proof.

We can then prove full implementation in interim rationalizable coincides with full implementation given $E^I$.

**Remark 1.** *A SCF $f$ is implementable in interim rationalizable strategies if and only if it is implementable given $E^I$.*

We can also notice that the argument goes through even if we slightly tweak the definition of the operator $b$ by requiring $\lambda_i \in \Delta^{t_i}(T_{-i}, S_{-i}) \subseteq \Delta(T_{-i}, S_{-i})$, an approach similar to the one used in models of $\Delta$-rationalizability[13].

In particular, as $E^*$ and $E$ would still coincide, the resulting theory of expectations would still be BRC. Indeed, it is not even necessary for operator $b$ to reach a fixed point to get a BRC theory of expectations: it is enough to assume agents expect their opponents to perform at least one round of elimination of non-rationalizable strategies[14]. As an example, let:

$$\tilde{E}(\gamma) = \{e \in \mathcal{E}(\gamma) : \text{supp}(e_{t_i}^j) \in b_j(S_j)\}$$

So that each player $i$ does not expect her opponent to play dominated strategies. It is easy to see $\tilde{E}^*(\gamma) \neq \emptyset$: if that was not the case, those strategies would have been eliminated by operator $b$ and would not be contained in $b_j(S_j)$, thus reaching a contradiction.

[13] See for example Battigalli and Siniscalchi (2003) and Ollár and Penta (2017).

[14] A concept similar to k-rationalizability of Bernheim (1984), for $k = 2$.

### 4.2.3 Shared Expectations of Best Responding

Under rational expectations, all agents' share the same expectations about a given opponent $j$ and expect her to best respond to her expectations. Moreover, these expectations need to be *correct* for a profile to be an equilibrium. Consider now the following model of expectations, dropping this last correctness requirement:

$$E^S(\gamma) = \{e \in \mathcal{E}(\gamma) : e_i^j = e_{i'}^j, e_i^j(t_j) \in B_{t_j}(e_j), \text{ where } j, i, i' \in I\}$$

It is clear $E^S = E^{S*}$ by construction: therefore, as long as $E^S \neq \emptyset$, it satisfies BRC and thus WBRC. By Theorem 1 any function implementable given $E^S$ must be BIC, showing incentive compatibility is still necessary even if we drop the requirement agents' expectations are correct.

### 4.2.4 Expectations as Precedents

The framework described in this paper, with due modifications, could be applied to learning and evolutionary models as well. To provide a simple example with a crude model, suppose agents form expectations according to how the game has been played historically: any profile of strategies that has been a solution of the game in the past is a plausible expectation for today's players. For instance, we can imagine the mechanism has been set up at time $\chi = 0$ from the planner, and at each moment in time an identical set of $N$ agents comes in and plays it one-shot. Before playing, each agent is given access to a history $\{\sigma^\chi\}_{\chi \leq \bar{\chi}}$ of past solutions of the game. That is, $\sigma^\chi \in B(e)$ for $e \in E^\chi(\gamma)$, where for all $\gamma \in \Gamma$:

$$E^\chi(\gamma) = \{e \in \mathcal{E}(\gamma) : e_i = \sigma_{-i}^{\chi'}, \chi' < \chi\}$$

We can easily prove $E^{\chi^*}(\gamma) \neq \emptyset$ for all $\chi > 0$: therefore, every SCF $f$ that is implementable given $E^\chi$ will necessarily be SIRBIC as well. To prove so, consider $\sigma \in B(e)$ with $e \in E^\chi(\gamma)/E^0(\gamma)$. Then, for all $i \in I$ there exists $\chi_i < \chi$ such that $\sigma_i \in B_i(\sigma_{-i}^{\chi_i})$. Moreover, for all $j \neq i$, $\sigma_j^{\chi_i} \in B_j(\sigma_{-j}^{\chi_j})$ for some $\chi_j < \chi_i$. For each $i \in I$, we can then construct an expectation $e'$ such that $(\sigma_i, e_i) \in B(e')$ by setting $e_i' = \sigma_{-i}^{\chi_i}$ and $e_j' = \sigma_{-j}^{\chi_j}$ for $j \neq i$.

## 4.3 A comment about partial implementation

Even if not the main focus of the present paper, the framework presented here can provide some insights about partial implementation as well, complementing and extending the results of Kneeland (2022) for level-k models. In particular I find that, in order for BIC to be necessary in a partial implementation framework, one solution of the implementing mechanism has to *almost* be a Bayesian Nash Equilibrium.

Let a function $f$ be *partially implementable* whenever there exists a mechanism $\Gamma$ and $e \in E(\gamma)$ such that $\mu \circ \sigma = f$ for $\sigma \in B(e)$, dropping the uniqueness requirement we were imposing in our previous definition of full implementability. Partial implementation is usually regarded as yielding more permissive results than full implementation, as it does away with the conditions required to ensure all unwanted solutions of the game are knocked off. We show this is the case even in our setup: dropping the requirement of full implementation seems to enlarge considerably the set of solution concepts that allow implementation of non-BIC social choice functions.

Let me now consider the following properties.

**Definition 5** (Partial BRC)**.** *We say a model of expectations $E$ is PBRC for a mechanism $\gamma$ whenever there exists $e \in E(\gamma)$ and $\sigma \in B(e)$ such that for all $i \in I$, $t_i, t_i' \in T_i$:*

$$\int_{t_{-i}} u_i(\mu(\sigma(t)), t)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(\mu(\sigma_i(t_i'), \sigma_{-i}(t_{-i})), t)\, dp(t_{-i}|t_i)$$

*Moreover, if the inequality above holds for all $\sigma \in B(e)$ for $e \in E(\gamma)$, we say $E$ is strongly PBRC (SPBRC) for $\gamma$.*

So a model of expectations $E$ is PBRC whenever there is at least one profile of expectations $e$ and a solution $\sigma$ such that *for all* players and types it a best reply to play according to their true type rather than mimicking a different one. Notice this a stronger condition than WBRC discussed above, as WBRC allows for the solution chosen to be different for each player $i$ and pair of types $t_i, t_i'$. This entails that any solution that satisfies PBRC will almost be an equilibrium, highlighting again the tight relationship between BIC and BNE as a solution concept.

These properties allow us to (almost) characterize the set of all solution concepts that render BIC necessary[15].

**Theorem 2.** *If, whenever $f$ is implementable given $E$ via mechanism $\gamma$:*

- *$f$ is BIC, then $E$ is PBRC for $\gamma$*
- *$f$ is not BIC, then $E$ is not SPBRC for $\gamma$*

Theorem 2 highlights how BIC in a partial implementation framework seems to be a more "fragile" necessary condition than in a full implementation setup, as the set of models of expectations satisfying $PBRC$ is smaller than the set of those satisfying $WBRC$.

We can relate this to Example 2 in de Clippel et al. (2019), as it is easy to check a solution of the mechanism in that setup satisfies the condition above. From the discussion in the paper, we know reporting their own type and $f$ is a best reply for all players. Thus $\sigma(t_i) = \{t_i, f\}$ yields higher expected utility than $\sigma(t_i') = \{t_i', f\}$ for all $i \in I$ and $t_i, t_i' \in T_i$, which allows us to conclude the $f$ is indeed partially implementable.

# 5 Necessity of Bayesian Monotonicity

In this section, I will provide preliminary results on necessity of a monotonicity-like condition for full implementation, which I call Weak Interim Expectations Monotonicity (WIEM)[16].

Let a deception $\beta$ be a profile of correspondences $\beta_i : T_i \to 2^{T_i}/\emptyset$ such that $t_i \in \beta_i(t_i)$ for all $i \in I$ and $t_i \in T_i$. This approach allows agents to expect

[15]It is evident from the proof of Theorem 2 that the gap between PBRC and SPBRC to achieve a full characterization of the class of solution concepts that make BIC necessary for partial implementation. However, I avoid doing so as that would require the statement of the definition to depend on the planner's choice of $f$.

[16]I have not achieved a full characterization yet, I just have a set of conditions on the solution concept and deceptions that imply a monotonicity-like condition is necessary for implementation.

their opponents are mixing between their actions (cf. Serrano and Vohra (2010), Kunimoto (2019) and Kunimoto et al. (2020)).

As I will model WIEM on Kunimoto et al. (2020)'s Weak Interim Rationalizable Monotonicity (WIRM), let me define the following property to keep notation in the rest of the section lighter.

**Definition 6** (Deception-Compatible Distribution)**.** *We say a probability distribution $\psi_i \in \Delta(T_{-i} \times T)$ is $\beta_i(t_i)$-compatible whenever:*

1. $\psi_i(t_{-i}, \tilde{t}) > 0 \implies \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$
2. $marg_{T_{-i}} \psi_i(t_{-i}, \tilde{t}) = p_i(t_{-i}|t_i)$

*We moreover say a collection $\psi = \{\psi_i\}_{i \in I}$ is $\beta(t)$-compatible whenever it is $\beta_i(t_i)$-compatible for all $i \in I$.*

It is easy to notice deception compatibility subsumes the two properties $\psi_i$ is required to have in Kunimoto et al. (2020)'s definition of WIRM. We can interpret $\psi_i$ as a belief: $\psi_i(t_{-i}, \tilde{t})$ is then the probability player $i$ associates to her opponents of type $t_{-i}$ mimicking types $\tilde{t}_{-i}$ in a way that is compatible with deception $\beta_{-i}$. As in Kunimoto et al. (2020), I include $T_i$ in the domain of $\psi_i$ to account for the fact agents' expectations are allowed to depend on her type.

Let me now define:

**Definition 7** (Unacceptable Deception)**.** *A deception $\beta_i : T_i \to 2^{T_i}/\emptyset$ is said to be unacceptable whenever there exists $t' \in \beta(t)$ such that $f(t') \neq f(t)$.*

That is, a deception is unacceptable only it affects the planner's ability to implement the SCF of interest.

Let moreover $Y_i[t_i, f]$ denote the set of functions $y : T_{-i} \to A$ such that either $y(t) = f(t)$ for all $t_{-i} \in T_{-i}$ or:

$$\int_{t_{-i}} u_i(f(t), t)\, dp(t_{-i}|t_i) > \int_{t_{-i}} u_i(y(t), t)\, dp(t_{-i}|t_i)$$

As in Kunimoto et al. (2020), we can interpret such a set as the set of functions that are equal or strictly worse than $f$ for type $t_i$.

Lastly, let me say $f$ is not responsive to $i$ changing her type from $t_i$ to $t_i'$ (denoted as $t_i' \not\sim^f t_i$) whenever $f(t_i, t_{-i}) = f(t_i', t_{-i})$ for all $t_{-i} \in T_{-i}$. If $f$ is not responsive for all $t_i, t_i' \in T_i$, we moreover say $f$ is not responsive to agent's $i$ type.

We are now ready to define weak refutability.

**Definition 8** (Weakly Refutable Deception)**.** *We say deception $\beta$ that is unacceptable for an SCF $f$ is weakly refutable if there exists $i \in I$, $t_i \in T_i$ and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \not\sim^f t_i$ such that for all $\beta_i(t_i)$-compatible distributions $\psi_i \in \Delta(T_{-i} \times T)$ there exists an SCF $f'$ such that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$ and:*

$$\int_{t_{-i},\tilde{t}} u_i(f'(\tilde{t}), t)\, d\psi_i(t_{-i}, \tilde{t}) > \int_{t_{-i},\tilde{t}} u_i(f(t_i', \tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t})$$

The basic intuition for this definition hsa been provided by Jackson (1991) and Kunimoto (2019): to detect deceptions in a direct mechanism, the planner must find a way of knocking off unwanted untruthful equilibria without doing the same for truthful ones: this is precisely the role of $f'$ in our definition of WIEM. The main difference with the BNE framework is that function $f'$ needs not to be the same for all $\beta_i(t_i)$-compatible beliefs $\psi_i$: as a matter of fact, WIEM ensures there exists at least one such function $f'$ for each $\beta_i(t_i)$-compatible $\psi_i$. This multiplicity arises for two main reasons: as we relax the belief consistency requirement imposed by BNE and as we allow different types of player $i$ to hold different expectations about her opponents. I refer the reader to Kunimoto et al. (2020) for a more detailed discussion about the interpretations of this condition.

We are now ready to turn to the main restriction we will impose on the solution concept in order to make WIEM necessary. Let us define for any given $e \in E^*(\gamma)$:

$$\bar{e}_{t_i}(t_{-i}) = \int_{\tilde{t}} e_{\tilde{t}_i}(\tilde{t}_{-i})\, d\psi_i(t_{-i}, \tilde{t})$$

We can interpret this object as the profile of strategies that type $t_i$ of agent $i$ would expect if her opponents of types $t_{-i}$ were mimicking $\tilde{t}_{-i}$ according to distribution $\psi_i$: a sort of "expected deception".

**Definition 9** (Non Separability (NS))**.** *We say a deception $\beta$ satisfies No Separability for mechanism $\gamma \in \Gamma$ whenever for all $t \in T$, $\beta(t)$-compatible $\psi$ and $t' \in T$*

*there exists* $e \in E^*(\gamma)$, $s \in \Delta(S)$ *such that:*

$$s \in B_t(\bar{e}_t) \cap B_{t'}(e_{t'}) \implies s \in B_t(e'_t) \text{ for } e' \in E(\gamma)$$

*We say a deception* $\beta$ *is non-separable when it satisfies non-separability for all* $\gamma \in \Gamma$.

That is, if we suppose $s$ is a solution to the game when the true type profile is $t' \in \beta(t)$ and $s_i$ a best reply for all $i \in I$ to the "expected deception" $\bar{e}_{t_i}$, then it is a solution to the game when the true type profile is actually $t$. To see why such a condition may lead to the necessity of BM-like conditions, consider their usual interpretation: the planner needs to give some type $t_i$ an incentive to uncover the deception other agents are playing, so that unwanted outcomes can be knocked off as equilibria or solutions to the mechanism. If there was no profile $s$ such that we simultaneously have $s \in B_{t'}(e_{t'})$ and $s \in B_t(e'_t)$, the planner would have no need to do so as there is no solution to the mechanism in which the planner is unable to tell apart (to separate) type profiles $t$ and $t'$.

To clarify why we weaken the property by requiring $s \in B_t(\bar{e}_t)$ as well, suppose instead $s \in B_{t'}(e_{t'})$ but $s \notin B_t(\bar{e}_t)$. Then, at least one player $i$ of type $t_i$ would have no incentive to play along with the deception if she expects the other to do so: the planner would then have a "natural ally" to defeat the unacceptable deception. Notice also that if a NS deception exists, then $E^*(\gamma) \neq \emptyset$ for all $\gamma \in \Gamma$: then, it follows $E$ is BRC. We have now the main definition for the section:

**Definition 10** (Weak Interim Expectations Monotonicity (WIEM))**.** *We say a SCF* $f$ *satisfies WIEM whenever all unacceptable and non-separable deceptions are weakly refutable.*

Let us now consider the main result for this section.

**Theorem 3.** *If* $f$ *is implementable given* $E$*, then it satisfies WIEM.*

I relegate the proof of this result in the Appendix: it is worth noting, however, this result follows from a very similar argument to the one proposed in Kunimoto et al. (2020).

Results in this and the previous Section suggest that it may not be possible to significantly expand the class of implementable social choice function. However, a word of caution is in order as they are still far from a full characterization and as I have not fully investigated whether the models considered in Section 4.2 satisfy non-separability for all deceptions $\beta$.

## 5.1 Examples

### 5.1.1 ICR

We can show every deception is non-separable for every $\gamma \in \Gamma$ for the ICR model of Kunimoto et al. (2020). Indeed, their proof for necessity of Weak Interim Correlated Monotonicity (WIRM) provides most of the argument, so I refer the reader to their paper for some of the details I omit here.

Suppose, for each $i \in I$, there exists a collection $\lambda_i \in \Delta(T_{-i} \times S_{-i})$ defined as follows:

$$\lambda_i(t_{-i}, s_{-i}) = \int_{\tilde{t}} \sigma_{-i}^{\tilde{t}_i} \, d\psi_i(t_{-i}, \tilde{t})$$

Where $\sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[s_{-i}] > 0$ implies $s_{-i} \in S_{-i}^{\gamma(T)}(\tilde{t}_{-i})$. The authors prove then that if there exists $t' \in \beta(t)$ such that $s^{t'} \in S^{\gamma(T)}(t')$ and $s_i^{t'}$ is a best reply to belief $\lambda_i$ for all $i \in I$, then we have $s^{t'} \in S^{\gamma(T)}(t)$ as well.

These statements can be easily translated into the language of our model. Consider any $e^*$ such that for all $i \in I$ and $\tilde{t}_i \in T_i$, $e^*_{\tilde{t}_i} = \sigma_{-i}^{\tilde{t}_i}$. It is clear $e^* \in E^*(\gamma)$, as each $e^*_{\tilde{t}_i} : T_{-i} \to \Delta(S_{-i})$ and for each $j \neq i$ and $t_j \in T_j$ the definition of $\sigma_{-i}^{\tilde{t}_i}$ implies $\text{supp}(e^*_{\tilde{t}_i}(t_j)) = \text{supp}(\sigma_{-i}^{\tilde{t}_i}(t_j)) \subseteq R_j^{\gamma(T)}(t_j)$. Therefore, as $s_i^{t'}$ is a best reply for type $t_i$ to $\lambda_i(t_{-i}, s_{-i})$, setting $\bar{e}_{t_i}(t_{-i}) = \int_{\tilde{t}} e^*_{\tilde{t}_i} \, d\psi_i(t_{-i}, \tilde{t})$ yields $s_i^{t'} \in B_{t_i}(\bar{e}_{t_i})$ for all $i \in I$ as well.

As for $s^{t'} \in S^{\gamma(T)}(t)$, there exist again a collection $\{\tilde{\lambda}_i\}_{i \in I}$ such that $s_i^{t'}$ is a best reply for type $t_i$ to belief $\tilde{\lambda}_i$, $\text{marg}_{T_{-i}} \tilde{\lambda}_i = p_i(t_{-i}|t_i)$ and $\tilde{\lambda}_i(t_{-i}, s_{-i}) > 0$ implies $s_{-i} \in R^{\gamma}(t)$. If for all $i \in I$ and $(t_{-i}, s_{-i}) \in (T_{-i}, S_{-i})$ we let $e_i(t_{-i})[s_{-i}] = \tilde{\lambda}_i(t_{-i}, s_{-i})$, it is easy to see $s_{-i} \in \text{supp}(e_i(t_{-i}))$ implies $\tilde{\lambda}_i(t_{-i}, s_{-i}) > 0$, and by definition of $\tilde{\lambda}_i$ we have $s_{-i} \in R^{\gamma(T)}(t_{-i})$. It follows $e \in E(\gamma)$ and $s_i^t \in B_{t_i}(e_i)$,

completing the argument.

# 6 Final Remarks

In this paper, I characterize the set of solution concepts that make BIC necessary for implementation of social choice functions. These and other preliminary results on a monotonicity-like condition suggest that the limits of implementation of functions may not be pushed too far from the ones of Bayesian Implementation (a takeway similar to Crawford (2021), de Clippel et al. (2019) and Kunimoto et al. (2020)).

My framework is quite general, but it still presents three main caveats. The first one we should consider is that focusing on single-valued social choice may be one of the reasons my results are so restrictive. Kneeland (2022) raises this point in the context of level-k implementation, and it should be investigated whether such an insight can be generalized to a broader class of models of behavior.

A second caveat lies in my definition of a model of expectations, which implicitly assumes $e_i$ is independent of $\sigma_i$. This causes some loss of generality: as an example, consider an agent maximizing her own payoff over $\Sigma_i$ under the expectation that other players will try to minimize her expected payoff. That is:

$$e_i(\sigma_i) \in \arg\min_{\sigma'_{-i}} \int_{t_{-i}} u_i(\mu(\sigma_i, \sigma_{-i}), t)\, dp_j(t_{-j}|t_j)$$

This concept, similar to the one of maximin, cannot be captured by the setup I present, as the minimizing profile $e_i$ will depend on the strategy $\sigma_i$ considered for player $i$. An extension of the model in this direction is certainly possible, but I neglect it in the context of the present paper in order not to make its notation more cumbersome and its insights less clear.

Third, this paper is silent about sufficient conditions for implementation. This is partly due to the difficulties intrinsic in constructing an implementing mechanism that works for a host of different solution concepts. Some of these difficulties may be sidestepped by focusing on implementation via direct mechanisms, which would allow us to describe the restrictions the model of expectations $E$ has to sat-

isfy for the (direct) implementing mechanism in terms of the social choice function $f$ of interest.

As argued in the introduction, my paper is related to both robust and continuous implementation, and it would be surely fruitful to study further the connection between those approaches and mine. In particular, the Oury and Tercieux (2012) and de Clippel et al. (2021) require the mechanism to implement $f$ continuously, so that small perturbations in the planner's model lead to small perturbations in the outcome prescribed by the SCF. The same ideas could be used to study whether small perturbations in the solution concept used generate discontinuous changes in the set of implementable social choice functions. This would be extremely useful to identify those necessary (or, possibly, sufficient) conditions carrying deeper intuition that does not depend on the particular solution concept used for implementation.

# References


Battigalli, P. and Siniscalchi, M. (2003). Rationalization and incomplete information. *Advances in Theoretical Economics*, 3(1).

Bergemann, D. and Morris, S. (2005). Robust mechanism design. *Econometrica*, 73(6):1771–1813.

Bergemann, D. and Morris, S. (2011). Robust implementation in general mechanisms. *Games and Economic Behavior*, 71(2):261–281.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, 52(4):1007–1028.

Brusco, S. (1995). Perfect bayesian implementation. *Economic Theory*, 5(3):419–444.

Crawford, V. P. (2021). Efficient mechanisms for level-k bilateral trading. *Games and Economic Behavior*, 127:80–101.

de Clippel, G. (2014). Behavioral implementation. *American Economic Review*, 104(10):2975–3002.

de Clippel, G., Saran, R., and Serrano, R. (2019). Level-$k$ Mechanism Design. *The Review of Economic Studies*, 86(3):1207–1227.

de Clippel, G., Saran, R., and Serrano, R. (2021). Continuous level-k mechanism design. *Working Paper*.

Dekel, E., Fudenberg, D., and Morris, S. (2007). Interim correlated rationalizability. *Theoretical Economics*, 2:15–40.

Eyster, E. and Rabin, M. (2005). Cursed equilibrium. *Econometrica*, 73(5):1623–1672.

Jackson, M. O. (1991). Bayesian implementation. *Econometrica*, 59(2):461–477.

Kneeland, T. (2022). Mechanism design with level-k types: Theory and an application to bilateral trade. *Journal of Economic Theory*, 201:105421.

Kunimoto, T. (2019). Mixed bayesian implementation in general environments. *Journal of Mathematical Economics*, 82:247–263.

Kunimoto, T., Saran, R., and Serrano, R. (2020). Interim rationalizable implementation of functions. *Working Paper*.

Maskin, E. (1999). Nash equilibrium and welfare optimality. *The Review of Economic Studies*, 66(1):23–38.

Maskin, E. and Sjöström, T. (2002). Implementation theory. *Handbook of social Choice and Welfare*, 1:237–288.

Ollár, M. and Penta, A. (2017). Full implementation and belief restrictions. *American Economic Review*, 107(8):2243–77.

Oury, M. and Tercieux, O. (2012). Continuous implementation. *Econometrica*, 80(4):1605–1637.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.

Serrano, R. (2004). The theory of implementation of social choice rules. *SIAM Review*, 46(3):377–414.

Serrano, R. and Vohra, R. (2010). Multiplicity of mixed equilibria in mechanisms: A unified approach to exact and approximate implementation. *Journal of Mathematical Economics*, 46(5):775–785.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118.

# Appendix A Generalized Model

In the definition of full implementation above, I implicitly assume $\sigma$ is such that each agent is best responding to their expectations, then it is a solution to the mechanism: as we consider an implementation problem, the second item entails that $f = \mu \circ \sigma$. In order to allow my setup to capture a larger class of models and solutions concepts, I will relax this assumption by allowing the planner to impose more constraints on the solution concept considered (including, among others, BNE and its refinements).

To do so, I will specify a *solution* as a pair $(\sigma, e)$ where $\sigma \in B(e)$ and $e$ is a profile of expectations as defined above. Let now $\tilde{\mathcal{E}} = \cup_{\gamma\in\Gamma}\mathcal{E}(\gamma)$ and $\tilde{\Sigma} = \cup_{\gamma\in\Gamma}\Sigma(\gamma)$. We can replace correspondence $E$ with a *theory of (expectations-based) behavior*. This will be formally defined as any correspondence $L : \Gamma \times T \to 2^{\tilde{\Sigma}\times\tilde{\mathcal{E}}}$ such that:

$$L(\gamma, t) \subseteq \{(\sigma(t), e_t) \in \Sigma(\gamma) \times \mathcal{E}(\gamma) : \sigma_i(t_i) \in B_{t_i}(e_{t_i}), \text{ for all } i \in I\}$$

We will interpret $L(\gamma, t)$ as the set of profiles $\sigma(t)$ that constitute a solution to mechanism $\gamma$ when the type profile is $t$, together with a profile of expectations $e_t$ sustaining it. This augments the "classical" notion of solution concept by explicitly specifying what expectations each agent is best responding to. This is not a new insight: for example, a complete description of a Perfect Bayesian equilibrium requires specifying both the strategy profiles played and the beliefs sustaining them. To simplify notation, let me say $(\sigma, e) \in L(\gamma)$ whenever $(\sigma(t), e_t) \in L(\gamma, t)$ for all $t \in T$.

Notice that $L$ generalizes the concept of theory of expectations and the associated solution concept for the mechanism: as a matter of fact, for an arbitrary theory of expectations $E$ we can then define the correspondence:

$$L(\gamma) = \{(\sigma, e) \in \Sigma(\gamma) \times \mathcal{E}(\gamma) : e \in E(\gamma), \sigma \in B(e)\}$$

Which will contain all pairs $(\sigma, e)$ such that $e \in E(\gamma)$ and such that $\sigma$ is a best reply to $e$ for all players. We can then generalize our definition of full implementability as follows.

**Definition 11.** *We say a SCF $f$ is fully implementable given $L$ whenever there exists $\gamma$ such that:*

- $L(\gamma) \neq \emptyset$
- $f = \mu \circ \sigma$ *for all profiles* $\sigma$ *such that* $(\sigma, e) \in L(\gamma)$

As before, let me refer to "full implementation" as "implementation" for the remainder of the section. In order to extend my main result, let me moreover substitute WBRC with:

**Definition 12** (Weak Theory of Behavior Consistency (WTBC))**.** *We say a model of expectations $E$ satisfies WBTC whenever for all $i \in I$, $t_i, t_i' \in T_i$ there exists $(\sigma, e) \in L(\gamma)$ such that:*

$$\int_{t_{-i}} u_i(\mu(\sigma(t), t)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(\mu(\sigma(t_i', t_{-i}), t)\, dp(t_{-i}|t_i)$$

As before, this property follows immediately if there exists a BNE $\sigma$ such that $(\sigma, e) \in L(\gamma)$. We can also provide a sufficient condition in the spirit of BRC. Similarly to our definition for $E^*$, let:

$$L^*(\gamma) = \{(\sigma, e) \in L(\gamma) : ((\sigma_i, e_{t_i}), e') \in L(\gamma), \text{ for all } t_i \in T_i, i \in I\}$$

The interpretation is the same we provided above for $E^*$: $L^*(\gamma)$ collects all strategy and expectation profiles such that type $t_i$ can justify her strategy by arguing the strategy profile $(\sigma_i, e_{t_i})$ she thinks will prevail in the mechanism will indeed be a solution to it for some supporting expectation profile $e'$.

**Definition 13** (Theory of behavior Consistency (TBC))**.** *We say $L$ is TBC whenever $L^*(\gamma) \neq \emptyset$ for all $\gamma \in \Gamma$.*

It is then easy to show any equilibrium theory of behavior is actually TBC for all mechanism $\gamma$ such that $L(\gamma) \neq \emptyset$. As for each $\gamma \in \Gamma$ there exists $(\sigma, e) \in L(\gamma)$ such that $\sigma \in B(e)$ and $e = (\sigma_{-i})_{i \in I}$, it is enough to let $e' = e = (\sigma_{-i})_{i \in I}$: it is clear then that $((\sigma_i, e_{t_i}), e') = (\sigma, e) \in L(\gamma)$.

**Theorem 4.** *Suppose $f$ is implementable given $L$ via mechanism $\gamma$. Then $f$ is BIC if and only if $E$ is WTBC for $\gamma$.*

By the same argument of Theorem 1, it moreover follows easily that:

**Corollary 1.** *Suppose $f$ is BIC is implementable given $E$ via mechanism $\gamma$. If $(\sigma, e) \in L(\gamma)$ implies $(\sigma', e) \in L(\gamma)$ for all $\sigma' \in B(e)$ and $e \in E(\gamma)$, then $f$ is SIRBIC.*

That is, if all best replies to a profile of expectations are solutions to the mechanism (as it is the case for the examples discussed in Section 4.2), then SIRBIC obtains for free from BIC and implementability given $E$. This suggests that the strengthening of BIC to SIRBIC in de Clippel et al. (2019) and Kunimoto et al. (2020) is a byproduct of the assumption all best replies to agents' expectations are solution to the mechanism rather than the use of a non-equilibrium model of behavior.

## A.1 Examples

### A.1.1 Nash equilibrium and refinements

The setup proposed can capture (Bayes)-Nash equilibrium if we impose:

$$L^{BN}(\gamma) = \{(\sigma, e) : \sigma \in B(e), e_i = \sigma_{-i}\}$$

It is then clear the set of Nash equilibria corresponds with the set of $\sigma \in \Sigma$ such that $(\sigma, e) \in L^{BN}(\gamma)$. Notice also that, as all solutions in $L^{BN}(\gamma)$ are BNEs, $L^{BN}$ is WTBC. This is still true for any refinement of BNE as well, implying any $L \subseteq L^{BN}$ is WTBC. These findings are consistent with the literature on dynamic mechanism design: for example, for the case of Perfect Bayesian Equilibrium, Brusco (1995) finds that BIC is necessary for implementation.

### A.1.2 Cursed equilibrium

This setup can capture the "Cursed Equilibrium" solution concept from Eyster and Rabin (2005). This leads to the following theory of behavior:

$$L^{CE}(\gamma) = \{(\sigma, e) : \sigma \in B(e), e_{t_i} = (1 - \chi)\sigma_{-i} + \chi\bar{\sigma}_{-i}(t_i)\}$$

Where:

$$\bar{\sigma}_{-i}(t_i) = \int_{t_{-i}} \sigma_{-i}(t_{-i})\, dp(t_{-i}|t_i)$$

It is possible to prove $L^{CE}$ is WTBC for all mechanisms $\gamma$ if agents have private values, but it is not otherwise.

To prove the argument for private values, suppose $(\sigma, e) \in L^{CE}(\gamma)$. As such, we have for $t_i \in T_i$ and $s_i \in \Delta(S_i)$:

$$(1-\chi)\int_{t_{-i}} u_i((\sigma(t_i), \sigma_{-i}(t_i)), t_i)\, dp(t_{-i}|t_i) + \chi \int_{t_{-i}} u_i((\sigma(t_i), \bar{\sigma}_{-i}(t_i)), t_i)\, dp(t_{-i}|t_i) \geq$$
$$(1-\chi)\int_{t_{-i}} u_i(s_i, \sigma_{-i}(t_i)), t_i)\, dp(t_{-i}|t_i) + \chi \int_{t_{-i}} u_i(s_i, \bar{\sigma}_{-i}(t_i)), t_i)\, dp(t_{-i}|t_i)$$

Then, as $u_i(\cdot)$ does not depend on $t_{-i}$, it follows by linearity of expected utility:

$$\int_{t_{-i}} u_i((\sigma(t), t_i)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(s_i, \sigma_{-i}(t_{-i}), t_i)\, dp(t_{-i}|t_i)$$

Which, as $(\sigma, e) \in L^{CE}(\gamma)$, concludes the proof.

However, the same is not true for all mechanisms $\gamma$ if an agent's payoff depends on the type of her opponents. For example, for any given value of $\chi \in (0, 1]$, we can construct the following two-player game:

| | | Player $C$ | |
|---|---|---|---|
| | | $A$ | $B$ |
| Player $R$ | $A$ | $t_R, t_C$ | $t_R + \zeta t_C, 0$ |
| | $B$ | $0, t_C + \zeta t_R$ | $0, 0$ |

Where $t_i \in \{-1, 1\}$ for $i \in \{R, C\}$, each type profile happens with equal probability and $\zeta \in (2, \frac{2}{1-\chi})$[17]. It can be easily determined that the only cursed equilibrium of this game is for type 1 to play $A$ and for type $-1$ to play $B$. To prove so, consider any solution $\sigma$ of the game. Then:

$$\bar{\sigma}_i(t_i)[A] = \frac{1}{2}\sigma_j(1)[A] + \frac{1}{2}\sigma_j(-1)[A]$$

It is clear the payoff of $B$ is always 0 for either player. As for action $A$, let me first calculate the "rational" part of the payoff of type $t_i$ of player $i$ as follows:

$$\int_{t_{-i}} u_i((A, \sigma_{-i}(t_{-i})), t_i)\, dp(t_{-i}|t_i) =$$

$$\frac{1}{2}[\sigma_{-i}(1)[A]t_i + (1 - \sigma_{-i}(1)[A])(t_i + \zeta)] + \frac{1}{2}[\sigma_{-i}(-1)[A]t_i + (1 - \sigma_{-i}(-1)[A])(t_i - \zeta)] =$$

$$t_i - \frac{1}{2}\zeta(\sigma_{-i}(1)[A] - \sigma_{-i}(-1)[A])$$

While the "cursed" part of the payoff is:

$$\int_{t_{-i}} u_i((A, \bar{\sigma}_{-i}(t_i)), t_i)\, dp(t_{-i}|t_i) =$$

$$\bar{\sigma}_i(t_i)[A]t_i + (1 - \bar{\sigma}_i(t_i)[A])[\frac{1}{2}(t_i + \zeta) + \frac{1}{2}(t_i - \zeta)] = t_i$$

Therefore, total payoff from playing $A$ is:

$$t_i - \frac{1}{2}(1 - \chi)\zeta(\sigma_{-i}(1)[A] - \sigma_{-i}(-1)[A])$$

To ensure type 1 will play $A$ with probability 1, it is enough to ensure:

$$1 - \frac{1}{2}(1 - \chi)\zeta > 0 \iff \zeta < \frac{2}{1 - \chi}$$

[17] In the discussion below, let me focus on the case $\chi < 1$. The case for $\chi = 1$ follows easily from the same steps as long as $\zeta > 2$.

While type $t_i = -1$ will play $B$ with probability 1 if:

$$-1 - \frac{1}{2}(1-\chi)\zeta < 0 \iff \zeta > \frac{-2}{1-\chi}$$

Which holds true as we assume $\zeta > 2$.

Thus, for both agents $i$, $\sigma_i(1)[A] = 1$ and $\sigma_i(-1)[A] = 0$. Notice moreover this is the only solution to the game, and that it does not satisfy WTBC. For example, for type $t_i = 1$ of player $i$, $\zeta > 2$ implies:

$$\frac{1}{2}t_i + \frac{1}{2}(t_i - \zeta) = 1 - \frac{1}{2}\zeta < 0$$

Thus type $t_i = 1$ would have liked, if she were not ignoring the correlation between her opponents' strategies and types, to mimic type $t_i = -1$: this suggests it would in principle be possible to implement non-BIC functions using cursed equilibrium as a solution concept for the case of common values.

### A.1.3 Expectation-dependent utility

The setup presented here does not rely on utility being independent of $i$'s expectations or expectations of her opponents. For example, consider the following adaptation of the model of fairness equilibrium from Rabin (1993) to games of incomplete information. To keep this model comparable with the one from the author, let $N = 2$. We will moreover denote as $\pi_i(\mu(\sigma), t)$ the "material payoff" $i$ derives from the outcome associated to profile $\sigma$, $i$'s beliefs on how kind $j$ is being to her as $\tilde{f}_j(e_i^j, e_j^i)$ and player $i$'s kindness towards $j$ as $f_i(\sigma_i, e_i^j)$. We can then write $i$'s utility function as:

$$u_i(\sigma, t, e) = \pi(\mu(\sigma), t) + \tilde{f}_j(e_i^j, e_j^i, t)\,[1 + f_i(\sigma_i, e_i, t)]$$

The set of equilibria in the model can then be captured by the following theory of behavior:

$$L^{ED}(\gamma) = \left\{ (\sigma, e) : e_i^j = \sigma_{-j}, \sigma_i \in \arg\max_{\sigma_i' \in \Sigma_i} \int_{t_j} u_i((\sigma_i', \sigma_j), t, e)\, dp(t_j|t_i) \right\}$$

This entails rational expectations hold, and so that $L^{ED}$ is WTBC.

# Appendix B Proofs

*Proof of Theorem 1.* Let $f$ be implementable given $E$ via mechanism $\gamma$, and let $E$ be WBRC for $\gamma$. Consider now arbitrary $i \in I$, $t_i, t_i' \in T_i$. As $E$ satisfies WBRC, there exists $e \in E(\gamma)$ and $\sigma \in B(e)$ such that:

$$\int_{t_{-i}} u_i(\mu(\sigma(t), t)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(\mu(\sigma(t_i', t_{-i}), t)\, dp(t_{-i}|t_i)$$

As $\sigma \in B(e)$ and $e \in E(\gamma)$, implementability of $f$ yields $\mu(\sigma(t)) = f(t)$ and $\mu(\sigma(t_i', t_{-i})) = f(t_i', t_{-i})$ for all $t_{-i} \in T_{-i}$. Therefore, for $i \in I$ and $t_i, t_i' \in T_i$:

$$\int_{t_{-i}} u_i(f(t), t)\, dp_i(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(f(t_i', t_{-i}), t)\, dp_i(t_{-i}|t_i)$$

Which is enough to prove BIC holds as our choice of $t$ and $t_i'$ was arbitrary.

To prove we can strengthen it to SIRBIC, let me proceed by contradiction and suppose that indeed the incentive constraint above holds with equality. Define then $\tau_i : T_i \to \Sigma_i$ as agreeing with $\tilde{\sigma}_i$ except for the fact $\tau_i(t_i) = \sigma_i(t_i')$. As $\sigma$ is a solution to the mechanism, there exist $e$ such that $\sigma \in B(e)$. Moreover, as $\tau$ yields the same expected utility as $\sigma$ conditional on expectations $e$, $\sigma \in B(e)$ implies $\tau \in B(e)$. Then by the definition of implementation above, we have for all$t_{-i} \in T_{-i}$:

$$f(t_i, t_{-i}) = \mu(\tau(t_i), \sigma_{-i}(t_{-i})) = \mu(\sigma(t_i', t_{-i})) = f(t_i', t_{-i})$$

Which concludes the proof for the first statement.

As for the second one, suppose $E$ is not WBRC for $\gamma$: then for some $i \in I$ and $t_i, t_i' \in T_i$ there exist no $e$ and $\sigma \in B(e)$ such that:

$$\int_{t_{-i}} u_i(\mu(\sigma(t), t)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(\mu(\sigma(t_i', t_{-i}), t)\, dp(t_{-i}|t_i)$$

That is, for all $e \in E(\gamma)$ and $\sigma \in B(e)$:

$$\int_{t_{-i}} u_i(\mu(\sigma(t_i', t_{-i}), t)\, dp(t_{-i}|t_i) > \int_{t_{-i}} u_i(\mu(\sigma(t), t)\, dp(t_{-i}|t_i)$$

As $\sigma \in B(e)$ and $f$ is implementable, it follows:

$$\int_{t_{-i}} u_i(f(t_i', t_i), t)\, dp(t_{-i}|t_i) > \int_{t_{-i}} u_i(f(t), t)\, dp(t_{-i}|t_i)$$

Therefore, either $f$ is either not implementable or not BIC. As SIRBIC implies BIC, this leads to a contradiction with our premises, concluding the proof. □

*Proof of Remark 1.* Let me denote:

$$B(E^I, \gamma) = \{\sigma \in \Sigma : \sigma \in B(e), e \in E^I\}$$

And as $\mu(B(E^I, \gamma))$ the corresponding set of functions mapping type profiles into lotteries over outcomes. Notice now that for any $\gamma \in \Gamma$, $\mu(R^{\gamma(T)}) \subseteq \mu(B(E^I, \gamma))$. As $s \in R^{\gamma(T)}$ implies that for each profile $t_{-i}$ there exists a probability distribution $\lambda_i \in \Delta(T_{-i}, S_{-i})$ over rationalizable actions such that $s_i \in B(\lambda_i)$ for all $i \in I$: it then follows that $e = (\lambda_i)_{i \in I} \in E^I$ and $\mu(s) \in \mu(B(E^I, \gamma))$.

If $f$ is implementable over $E^I$, there exists $\gamma$ such that $\mu(B(E^I, \gamma))$ is a singleton and as $R^{\gamma(T)} \neq \emptyset$ it follows $\mu(R^{\gamma(T)}) = \mu(B(e)) = f$, proving the only if direction. Suppose now $f$ is implementable in interim rationalizable strategies: then there exists a mechanism $\gamma$ such that for any rationalizable action profiles $s$ and $s'$, $\mu(s) = \mu(s') = f$. As any profile $s \in \text{supp}(\sigma)$ is rationalizable and thus belongs to $R^{\gamma(T)}$, this entails $\mu(\sigma) = \mu(s) = f$ for all $\sigma \in B(E^I)$ and $s \in R^{\gamma(T)}$. This concludes the proof. □

*Proof of Theorem 2.* Suppose BIC $f$ is weakly implementable given $E$. Then, for all $\sigma$ such that $f = \mu \circ \sigma$, $i \in I$ and $t_i, t_i' \in T_i$ we have:

$$\int_{t_{-i}} u_i(\mu(\sigma(t)), t)\, dp(t_{-i}|t_i) =$$
$$\int_{t_{-i}} u_i(f(t), t)\, dp(t_{-i}|t_i) \geq$$
$$\int_{t_{-i}} u_i(f(t_i', t_i), t)\, dp(t_{-i}|t_i) =$$
$$\int_{t_{-i}} u_i(\mu(\sigma_i(t_i'), \sigma_{-i}(t_{-i})), t)\, dp(t_{-i}|t_i)$$

As by implementability there exists at least one solution $\sigma$ such that $\sigma \in B(e)$ for $e \in E(\gamma)$ and $f = \mu \circ \sigma$, this is enough to prove the statement.

Suppose now $f$ is weakly implementable given $E$ and that for all $\sigma \in B(e)$

such that $e \in E(\gamma)$, $i \in I$ and $t_i, t_i' \in T_i$ we have:

$$\int_{t_{-i}} u_i(\mu(\sigma(t)), t)\, d \geq \int_{t_{-i}} u_i(\mu(\sigma_i(t_i'), \sigma_{-i}(t_{-i})), t)\, dp(t_{-i}|t_i)$$

The result then follows immediately by the fact there exists at least one such solution with $f = \mu \circ \sigma$, concluding the proof. □

*Proof of Theorem 3.* If there is no deception that is both unacceptable and non-separable, WIEM holds vacuously. Suppose then that there exists a unacceptable and non-separable deception $\beta$: by the argument above, we then know $E$ is BRC.

Let me then state the following Lemma:

**Lemma 1.** *Suppose $f$ is implementable given BRC $E$ via mechanism $\gamma$, and that there exist a deception $\beta$ that is not weakly refutable. Then for all $i \in I$, $t \in T$ and $t' \in \beta(t)$, there exist a $\beta(t)$-compatible collection $\psi = \{\psi_i\}_{i \in I}$ such that for all $s_i \in S_i$ and $e \in E^*(\gamma)$:*

$$\int_{t_{-i}, \tilde{t}} u_i(f(t_i', \tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) \geq \int_{t_{-i}, \tilde{t}} u_i(\mu(s_i, e_{\tilde{t}_i}(\tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t})$$

The proof of this result is basically the same as in Kunimoto et al. (2020), and it is provided separately in this Appendix.

Suppose now, for the sake of contradiction, that $f$ is implementable and it does not satisfy WIEM. Then there exists an unacceptable and non-separable deception $\beta$ that is not weakly refutable. As $\beta$ is also NS, we know for all $t \in T$ and $\beta(t)$-compatible $\psi$ there exist $e \in E^*(\gamma)$, $t' \in \beta(t)$ and $s \in \Delta(S)$ such that $s \in B_t(\bar{e}_t) \cap B_{t'}(e_{t'})$ implies $s \in B_t(e_t')$ for $e' \in E(\gamma)$.

Consider then such $e$. By Lemma 1, there exists a $\beta(t)$-compatible $\psi$ such for all $i \in I$ we have:

$$\int_{t_{-i}, \tilde{t}} u_i(f(t_i', \tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) \geq \int_{t_{-i}, \tilde{t}} u_i(\mu(s_i, e_{\tilde{t}_i}(\tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t})$$

By linearity of expected utility, we can rewrite the RHS as follows:

$$\int_{t_{-i}, \tilde{t}} u_i(\mu(s_i, e_{\tilde{t}_i}(\tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) = \int_{t_{-i}} u_i(\mu(s_i, \bar{e}_{t_i}(t_{-i}), t)\, dp(t_{-i}|t_i)$$

As $e \in E^*(\gamma)$, we moreover know $\sigma \in B(e)$ implies that $\mu(\sigma_i(t'_i), e_{\tilde{t}_i}(\tilde{t}_{-i})) = f(t'_i, \tilde{t}_{-i})$ for all $\tilde{t}_{-i} \in T_{-i}$. We can then rewrite the LHS as follows:

$$\int_{t_{-i},\tilde{t}} u_i(f(t'_i, \tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) =$$
$$\int_{t_{-i},\tilde{t}} u_i(\mu(\sigma_i(t'_i), e_{\tilde{t}_i}(\tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) =$$
$$\int_{t_{-i}} u_i(\mu(\sigma_i(t'_i), \bar{e}_{t_i}(t_{-i}), t)\, dp(t_{-i}|t_i)$$

Then, for all $i \in I$ and $s_i \in \Delta(S_i)$:

$$\int_{t_{-i}} u_i(\mu(\sigma_i(t'_i), \bar{e}_{t_i}(t_{-i}), t)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(\mu(s_i, \bar{e}_{t_i}(t_{-i}), t)\, dp(t_{-i}|t_i)$$

Proving that $\sigma(t') \in B_t(\bar{e}_t)$. Then, as $\sigma(t') \in B(e_{t'})$ by our choice of $\sigma$, NS implies there exist $e' \in E(\gamma)$ such that $\sigma(t') \in B_t(e'_t)$. Thus, $\mu(\sigma(t')) = f(t) = f(t')$: this contradicts $\beta$ being unacceptable, concluding the proof. □

*Proof of Lemma 1.* Suppose $f$ is implementable given $E$ using mechanism $\gamma$. As such, for any profile $e \in E^*(\gamma)$ there exists $\sigma$ with $\sigma_i(t_i) \in B_{t_i}(e_{t_i})$ for all $i \in I$ and $t_i \in T_i$, and $\mu(\sigma(t)) = f(t)$. Consider now $\sigma_i(t_i)$: as $e \in E^*(\gamma)$, we know there exists a profile $\tilde{e} \in E(\gamma)$ such that $(\sigma_i(t_i), e_{t_i}(t_{-i})) \in B(\tilde{e}_{t_i})$ for all $i \in I$ and thus for all $t_{-i} \in T_{-i}$:

$$y^{\sigma_i(t_i),t_i}(t_{-i}) = \mu(\sigma_i(t_i), e_{t_i}(t_{-i})) = f(t)$$

If follows then by implementability that for all $s_i \in \Delta(S_{-i})$:

$$\int_{t_{-i}} u_i(f(t), t)\, dp_i(t_{-i}|t_i) =$$
$$\int_{t_{-i}} u_i(y^{\sigma_i(t_i),t_i}(t_{-i}), t)\, dp_i(t_{-i}|t_i) =$$
$$\int_{t_{-i}} u_i(\mu(\sigma_i(t_i), e_{t_i}(t_{-i}), t)\, dp_i(t_{-i}|t_i) \geq$$
$$\int_{t_{-i}} u_i(\mu(s_i, e_{t_i}(t_{-i}), t)\, dp_i(t_{-i}|t_i) =$$
$$\int_{t_{-i}} u_i(y^{s_i,t_i}(t_{-i}), t)\, dp_i(t_{-i}|t_i)$$

We now show that if there exists $t_{-i} \in T_{-i}$ and $s_i$ such that $y^{s_i,t_i} \neq f(t)$, then the inequality above holds strictly. If that was not the case, $s_i \in B_{t_i}(e_{t_i})$, so that for all $t_{-i} \in T_{-i}$ we have $\mu(s_i, e_{t_i}(t_{-i})) = \mu(\sigma_i(t_i), e_{t_i}(t_{-i})) = f(t)$ as well. This yields a contradiction with $y^{s_i,t_i} \neq f(t)$, proving the inequality is indeed strict: this entails $y^{s_i,\tilde{t}_i} \in Y_i[\tilde{t}_i, f]$ f, which will come in handy later.

We now show there exist for all $i \in I$, $t \in T$ and $t' \in \beta(t)$ and a collection $\psi = \{\psi_i\}_{i \in I}$ such that, for all $i \in I$, $\psi_i$ is $\beta_i(t_i)$-compatible and for all $f'$ with $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ and $\tilde{t}_i \in T_i$:

$$\int_{t_{-i},\tilde{t}} u_i(f(t'_i, \tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) \geq \int_{t_{-i},\tilde{t}} u_i(f'(\tilde{t}), t)\, d\psi_i(t_{-i}, \tilde{t})$$

Consider arbitrary $t \in T$ and $t' \in \beta(t)$. We then construct function $\psi$ as follows. As for all $i \in I$ such that $t'_i \not\sim_i^f t_i$, we know by the fact $\beta$ is not weakly refutable that there exists at least one $\tilde{\psi}_i$ with the characteristics required: we then set $\psi_i = \tilde{\psi}_i$[18]. For all $i \in I$ such that $t'_i \sim_i^f t_i$, let $\psi_i(t_{-i}, \tilde{t}) = p_i(t_{-i}|t_i)$ whenever $\tilde{t}_i = t_i$ and $\tilde{t}_{-i} = t_{-i}$ and $\psi_i(t_{-i}, \tilde{t}) = 0$ otherwise. We can notice immediately that, for all $i \in I$, $\psi$ is satisfies $\tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ whenever $\psi_i(t_{-i}, \tilde{t}) > 0$ and $p_i(t_{-i}|t_i) = \int_{\tilde{t}} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$. Then for any $f'$ with $f'(\tilde{t}_i, f) \in Y_i(\tilde{t}_i, f)$ and all $\tilde{t}_i \in \tilde{T}_i$:

$$\begin{aligned}
&\int_{t_{-i},\tilde{t}} u_i(f(t'_i, \tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) = \\
&\int_{t_{-i},\tilde{t}} u_i(f(t_i, \tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) = \\
&\int_{t_{-i}} u_i(f(t), t)\, dp_i(t_{-i}|t_i) \geq \\
&\int_{t_{-i}} u_i(f'(t_i, t_{-i}), t)\, dp_i(t_{-i}|t_i) = \\
&\int_{t_{-i},\tilde{t}} u_i(f'(\tilde{t}), t)\, d\psi_i(t_{-i}, \tilde{t})
\end{aligned}$$

Where the first equality is due to the fact $t'_i \sim_i^f t'_i$, the second and fourth are by construction of $\psi_i$ and the inequality follows from the fact $f(t_i, \cdot) \in Y_i(t_i, f)$. To complete the proof, we have just to piece the two results we proved together

[18] If multiple such distributions exist, the choice of the one to be used is immaterial.

keeping in mind $y^{s_i,\tilde{t}_i} \in Y_i[t_i, f]$:

$$\int_{t_{-i},\tilde{t}} u_i(f(t_i', \tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t}) \geq$$
$$\int_{t_{-i},\tilde{t}} u_i(y^{s_i,\tilde{t}_i}(\tilde{t}_{-i})), t)\, d\psi_i(t_{-i}, \tilde{t}) =$$
$$\int_{t_{-i},\tilde{t}} u_i(\mu(s_i, e_{\tilde{t}_i}(\tilde{t}_{-i}), t)\, d\psi_i(t_{-i}, \tilde{t})$$

This concludes the proof. □

*Proof of Theorem 4.* We will start with the only if direction. As $f$ is implementable given $L$, we know there exists an implementing mechanism $\gamma = (\mu, S)$ such that $L(\gamma) \neq \emptyset$. Suppose now $L$ is not WTBC for $\gamma$. Then for some $i \in I$ and $t_i, t_i' \in T_i$ there exist no $(\sigma, e) \in L(\gamma)$ such that[19]:

$$\int_{t_{-i}} u_i(\mu(\sigma(t), t)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(\mu(\sigma(t_i', t_{-i}), t)\, dp(t_{-i}|t_i)$$

That is, for all $(\sigma, e) \in L(\gamma)$:

$$\int_{t_{-i}} u_i(\mu(\sigma(t_i', t_{-i}), t)\, dp(t_{-i}|t_i) > \int_{t_{-i}} u_i(\mu(\sigma(t), t)\, dp(t_{-i}|t_i)$$

As $f$ is implementable, it follows:

$$\int_{t_{-i}} u_i(f(t_i', t_i), t)\, dp(t_{-i}|t_i) > \int_{t_{-i}} u_i(f(t), t)\, dp(t_{-i}|t_i)$$

Therefore, either $f$ is either not implementable or not BIC: this leads to a contradiction with our premises, concluding this direction of the proof.

Let us now move to the converse. As $f$ is implementable given WTBC $L$, there exists an implementing mechanism $\gamma = (\mu, S)$ such that $L(\gamma) \neq \emptyset$. Consider arbitrary $i \in I$, $t_i, t_i' \in T_i$. As $f = \mu \circ \sigma$ for $(\sigma, e) \in L(\gamma)$, WTBC implies there exist $(\sigma, e) \in L(\gamma)$ such that:

$$\int_{t_{-i}} u_i(\mu(\sigma(t), t)\, dp(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(\mu(\sigma(t_i', t_{-i}), t)\, dp(t_{-i}|t_i)$$

[19] We know $L(\gamma) \neq \emptyset$ as $f$ is implementable given $L$.

As $(\sigma, e) \in L(\gamma)$, implementability of $f$ yields $\mu(\sigma(t)) = f(t)$ and $\mu(\sigma(t_i', t_{-i})) = f(t_i', t_{-i})$ for all $t_{-i} \in T_{-i}$. Therefore, for $i \in I$ and $t_i, t_i' \in T_i$:

$$\int_{t_{-i}} u_i(f(t), t)\, dp_i(t_{-i}|t_i) \geq \int_{t_{-i}} u_i(f(t_i', t_{-i}), t)\, dp_i(t_{-i}|t_i)$$

Which is enough to prove BIC holds as our choice of $t$ and $t_i'$ was arbitrary. $\square$