BROWN
Orlando Bravo Center
for Economic Research

Mechanism Design without Rational Expectations

Graduate Student Bravo Working Paper # 2023-001

Giacomo Rubbini*

May 5, 2023
Most recent version

Abstract

Does dropping the rational expectations assumption mean the social planner can implement a larger class of social choice rules? This paper proposes a generalized model of implementation that does not assume rational expectations and characterizes the class of solution concepts requiring Bayesian Incentive Compatibility for full implementation. Surprisingly, full implementation of social choice functions turns out not to be significantly more permissive than with rational expectations. This implies some classical results, such as the impossibility of efficient bilateral trade (Myerson and Satterthwaite, 1983) hold for a broad range of non-equilibrium solution concepts, confirming their relevance even in boundedly rational setups.

Keywords: Mechanism Design, Bounded Rationality, Rational Expectations

JEL: C72, D78, D82

*Department of Economics, Brown University, giacomo_rubbini@brown.edu. I am indebted to Roberto Serrano for his guidance and support. I wish to thank Tim Bergling, Pedro Dal Bó, Pietro Dall'Ara, Geoffroy De Clippel, Jack Fanning, Ricardo Fonseca, Takashi Kunimoto, Teddy Mekonnen, Zeky Murra Anton, Marco Petterson, Cosimo Petracchi, Rene Saran, Silvio Sorbera, Rajiv Vohra, and participants to various talks for useful comments and suggestions. All errors are my own.

Mechanism Design without Rational Expectations

Giacomo Rubbini*

May 5, 2023
Most recent version

Abstract

Does dropping the rational expectations assumption mean the social planner can implement a larger class of social choice rules? This paper proposes a generalized model of implementation that does not assume rational expectations and characterizes the class of solution concepts requiring Bayesian Incentive Compatibility for full implementation. Surprisingly, full implementation of social choice functions turns out not to be significantly more permissive than with rational expectations. This implies some classical results, such as the impossibility of efficient bilateral trade (Myerson and Satterthwaite, 1983) hold for a broad range of non-equilibrium solution concepts, confirming their relevance even in boundedly rational setups.

Keywords: Mechanism Design, Bounded Rationality, Rational Expectations

JEL: C72, D78, D82

*Department of Economics, Brown University, giacomo_rubbini@brown.edu. I am indebted to Roberto Serrano for his guidance and support. I wish to thank Tim Bergling, Pedro Dal Bò, Pietro Dall'Ara, Geoffroy De Clippel, Jack Fanning, Ricardo Fonseca, Takashi Kunimoto, Teddy Mekonnen, Zeky Murra Anton, Marco Petterson, Cosimo Petracchi, Rene Saran, Silvio Sorbera, Rajiv Vohra, and participants to various talks for useful comments and suggestions. All errors are my own.

1 Introduction

Can a 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? The answer to such a question depends on how this interaction pans out, and the mechanism design and implementation literature have extensively explored this problem using a variety of game-theoretic solution concepts.

Despite the popularity of Bayesian Nash Equilibrium (BNE) as a solution concept, insights from the experimental and behavioral literature have highlighted that equilibrium models may not provide a good prediction of agents' behavior in many settings. In these setups, the assumption that agents correctly anticipate their opponents' strategies (i.e., that agents have rational expectations) feels particularly unpalatable: for instance, when agents face a given interaction for the first time.

It remains unclear whether alternative solution concepts allow full implementation of a broader class of social choice rules than Bayesian Nash Equilibrium. Recent results about full implementation of social choice functions (SCF) in non-equilibrium solution concepts suggest that the answer to this question may be negative. For instance, de Clippel et al. (2019) and Kunimoto et al. (2020) prove that Bayesian Incentive Compatibility (BIC) is still necessary for full implementation of functions in level-k reasoning and interim correlated rationalizability. Results are instead more permissive for for full level-k implementation of social choice sets (SCS), for which BIC is no longer necessary (de Clippel et al., 2019).

This paper studies the limits of full implementation by characterizing the class of all solution concepts such that Bayesian Incentive Compatibility is necessary for full implementation. Our results suggest that we can generally not expect to expand the set of implementable social choice functions dramatically by moving to non-equilibrium solution concepts, while results about social choice sets are more permissive.

Our novel approach turns on its head implementation theory's standard approach of fixing a solution concept and only then deriving necessary conditions for full implementation, allowing us to search for a deeper property linking all solution concepts requiring BIC for full implementation. Other than providing useful guidance about the possibility of implementing non-BIC social choice rules, the results in this paper can allow us to extend some classical findings in the literature (for example, the impossibility theorem of Myerson and Satterthwaite (1983)) to a large class of solution concepts.

To achieve this goal, we propose a generalized model of full implementation that allows agents to hold arbitrary expectations about their opponents' strategies. This model allows

us to encompass all solution concepts in which agents best respond to their (possibly heterogeneous) expectations about their opponents. Our model nests the ones from Jackson (1991), de Clippel et al. (2019), Kunimoto et al. (2020), and Kneeland (2022) as special cases, unifying previous results about the necessity of Bayesian Incentive Compatibility for full implementation.

For the case of implementation of social choice functions, we show BIC is still necessary for implementation of functions if and only if the solution concept satisfies a novel property we call Weak Response Consistency (WRC). This property can be interpreted as requiring that, for each type of each agent, there exists a solution of the mechanism such that she does not have any incentive to mimic a different type. Differently from regular incentive compatibility, WRC does this solution is the same for all types of all agents[1]. Even if this property is not very restrictive, it is enough to establish necessity of BIC as full implementation of a function requires all the mechanism's solutions to yield the outcome prescribed by the social choice function.

Several solution concepts in the literature satisfy Weak Response Consistency (WRC): for instance, the level-k model of de Clippel et al. (2019), Interim Correlated Rationalizability (ICR, Kunimoto et al. 2020), and Bayesian Nash Equilibrium (BNE, Jackson 1991) satisfy this condition for any given mechanism as well. Notably, in the spirit of the so-called "Wilson Doctrine", WRC (and thus necessity of BIC) does not hinge on the assumption of common knowledge of rationality. The epistemic argument in Appendix B shows WRC is almost equivalent to requiring each type of each agent to know the type space and that she can mimic another type by inducing a different solution of the mechanism. Both requirements feel natural, confirming that WRC is a mild restriction and that the class of solution concepts satisfying WRC is rather large.

By providing a characterization of the set of solution concepts that allow implementation of BIC SCF, this paper also identifies which solution concepts allow for implementation of non-BIC social choice functions. Cursed equilibrium (CE, Eyster and Rabin 2005) and Naïve Bayesian Equilibrium (NBE, Gagnon-Bartsch et al. 2021) fall into the latter category, as they do not rule out the possibility that agents may not realize they could profitably mimic a different type. Existence of non-WRC solution concepts confirms that WRC has bite, and the characterization result hints at which solution concepts may be fruitful to investigate to study implementation of non-BIC social choice functions.

As for the problem of implementation of Social Choice Sets (SCS), WRC is not enough

[1]See discussion at the end of Section 2.

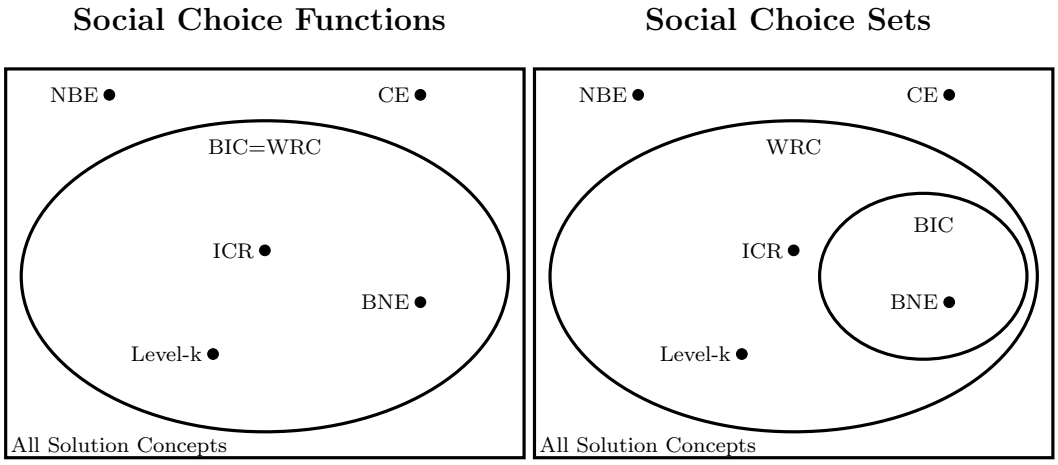

Figure 1: The class of solution concepts such that BIC is necessary for implementation coincides with the class of WRC solution concepts for full implementation of SCFs (left) and it is a subset of the class of WRC solution concepts for full implementation of SCSs (right).

to establish necessity of BIC for full implementation (Figure 1)[2]. Necessity of BIC for implementation of SCS turns out to be close to assuming rational expectations: this is then a relatively fragile result, unlikely to hold for most non-equilibrium models. WRC implies however that any implementable SCS must contain partially incentive compatible social choice functions: i.e., SCF that provide only some types and agents with the right incentives not to misrepresent their private information. This last result confirms and extends the findings of Kneeland (2022) for level-k reasoning models.

The contrast between the results for implementation of SCF and SCS suggests necessity of BIC is mainly driven by the requirement all solutions of the mechanism yield the same SCF when rational expectations do not hold. As agents understand all solutions will lead to the same outcome in the case of implementation of SCF, the same SCF must provide incentives to all agents not to misrepresent their type (that is, it must be BIC). If we allow different solutions to yield different outcomes instead (as in the case of full implementation of SCS), each type may believe a different solution of the mechanism (and the associated SCF) will obtain. The planner no longer needs the same SCF to simultaneously incentivize all types of all agents, unless rational expectations hold. In a sense, decoupling agents' expectations allows the planner to decouple the incentives she provides them. As rational expectations make this decoupling impossible, BIC is necessary for implementation in equilibrium solution concepts.

[2]The solution concept of de Clippel et al. (2019) is a case in point: even if their level-k reasoning model satisfies WRC, they show in Example 2 that it is possible to implement non-BIC SCS.

This discussion highlights a new tension behavioral mechanism design faces: while having a unique outcome for all solutions offers starker predictions in applications, it will often deliver restrictive results regarding incentive compatibility. This tension is absent in equilibrium solution concepts: regardless of the number of solution outcomes, rational expectations ensure BIC is necessary for implementation. This result follows again from the fact that both the uniqueness requirement and the rational expectations assumption do not allow the planner to decouple the incentives she provides to each agent from the ones she provides to other agents.

The importance of these results also stems from the fact they allow us to extend classical mechanism design findings to full implementation in any WRC solution concept. For the case of full implementation of functions, Section 6 considers three applications that extend to all WRC solution concepts the Revenue Equivalence Theorem (Myerson, 1981), impossibility of ex-post efficient and budget-balanced bilateral trade (Myerson and Satterthwaite, 1983) and of full surplus extraction in auctions. These applications highlight much of the underlying economic intuition for these results does indeed not hinge on the rational expectations assumption or the use of a particular equilibrium solution concept *per se*, and it remains central for the case of boundedly rational agents as well.

This paper advances the mechanism design and bounded rationality literature by providing a methodology to answer open questions about implementation with and without rational expectations. Unlike previous works, this paper investigates the robustness of necessity of BIC to changes in the solution concept. Previous papers focused instead on robustness in other model features: for example, Saran (2011) characterizes the domain of preferences on which the revelation principle holds, while Artemov et al. (2013) prove that the restrictiveness of robust virtual implementation stems from a particular zero-measure set of beliefs. This work also relates to approaches considering a planner with an inaccurate model of agents' payoffs and beliefs, and in particular to the literature about continuous (Oury and Tercieux, 2012) and robust (Bergemann and Morris, 2005) implementation. However, this paper focuses on a planner with an accurate model of payoffs and beliefs who is not sure how these map into the outcomes of strategic interaction and studies how sensitive restrictions on the set of implementable SCF (such as BIC) are to changes in this mapping.

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. As usual in the literature, incomplete

information is modeled by assuming there exists a set of types T_i for each agent $i \in I$, and that each agent knows her type but not the type of other players. Let $T = \times_{i \in I} T_i$ 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 $p_i(t_{-i}|t_i)$[3]. Assume that, for all $t \in T$, there exists at least one $i \in I$ such that t_{-i} belongs to the support of $p_i(\cdot|t_i)$ [4]. Preferences over lotteries have expected utility form, with Bernoulli utility $u_i : A \times T \to \mathbb{R}$. Abusing notation slightly, 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 Γ denote the set of all possible mechanisms the planner can design. Once the planner has committed to a mechanism, agents choose a strategy profile $\sigma : T \to \Delta(S)$. We will denote the set of such functions as Σ. For all $i \in I$, we will moreover let Σ_i (respectively, Σ_{-i}) denote the set of $\sigma_i : T_i \to \Delta(S_i)$ (respectively, $\sigma_{-i} : T_{-i} \to \Delta(S_{-i})$). For the rest of the paper, we will slightly abuse the notation above by considering $\mu(\sigma(t))$ to denote the lottery over A induced by $\sigma(t)$ under the outcome function μ.

A key feature of rational expectations models is that agents' expectations turn out to be correct in equilibrium: for example, if σ is a BNE, player i expects her opponents to play σ_{-i}. In order to relax the rational expectations assumption, we will consider a more general theory of agents' expectations. For a given mechanism γ, let $e_{i,t_i} \in \Sigma_{-i}$ represent the expectations of type t_i of agent i about her opponent. The set of all possible expectations over mechanism γ will be denoted as $\mathcal{E}(\gamma) = \times_{i \in I} \Sigma_{-i}$. As e_{i,t_i} is a strategy profile for players $j \neq i$, we will sometimes evaluate it at t_{-i}: thus, $e_{i,t_i}(t_{-i}) \in \Delta(S_{-i})$. To make notation more compact, define a mapping $e_i : T_i \to \Sigma_{-i}$ that assigns e_{i,t_i} to each type $t_i \in T_i$ and denote as e any profile $(e_i)_{i \in I} \in \mathcal{E}(\gamma)$.

The formulation above 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. Agents are also allowed to expect their opponents' actions to be correlated as $e_{i,t_i} \in \Sigma_{-i}$, and Σ_{-i} is not assumed to have a

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

[4]This assumption is not necessary for the argument, but it will make the notation more convenient as it will not be necessary to state results in terms of equivalent SCF.

product structure. This formulation makes it possible to accommodate models such as the Interim Correlated Rationalizability (ICR) model of Kunimoto et al. (2020)[5].

Let then a *theory of expectations* E be any correspondence mapping each mechanism γ 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 opponents to play a dominated strategy (the reader can refer to Section 5 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 expectation profiles the planner believes could happen with non-zero probability. This interpretation will reflect in the implementation concept used below, which requires the outcome prescribed by f to prevail regardless of the expectation profile considered.

Define a *theory of response* (or response correspondence) as any correspondence $R : E \times \gamma \to \Sigma$. A *solution concept* will then be a pair $\mathcal{S} = (R, E)$ consisting of both a theory of expectation formation and of how agents respond to these expectations. $\mathcal{S}$ maps each mechanism γ into a subset of Σ, which we can interpret the mechanisms's solutions[6]. Formally, let σ be a *solution* to a mechanism γ whenever $\sigma \in R(e)$ for $e \in E(\gamma)$.

For all $e \in E(\gamma)$, we will denote as $B_{i,t_i}(\sigma_{-i})$ the set of best replies for type t_i of i to the profile σ_{-i}[7]. That is, if $s_i \in B_{i,t_i}(\sigma_{-i})$ then for all $s_i' \in \Delta(S_i)$:

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

Let also $B(e)$ denote the set of $\sigma \in \Sigma$ such that for all $i \in I$ and $t \in T$, $s = \sigma(t)$ is such that $s_i \in B_{i,t_i}(e_{i,t_i})$.

We say a SCF f is *fully implementable* whenever there exists a mechanism γ that has at least one solution, and every such solution yields the outcome prescribed by f. Formally, we say a SCF f is fully implementable in $\mathcal{S}$ whenever there exists an implementing mechanism γ such that $\mu(\mathcal{S}(\gamma)) = f$ and $R(e) \neq \emptyset$ for all $e \in E(\gamma)$. Let moreover $\Gamma^f \subseteq \Gamma$ denote the class of all such mechanisms. Similarly, we say an SCS $F \neq \emptyset$ is fully implementable if there exists γ such that $\mu(\mathcal{S}(\gamma)) = F$ and $R(e) \neq \emptyset$ for all $e \in E(\gamma)$, and we denote the class of mechanisms implementing F as $\Gamma^F \subseteq \Gamma$.

[5]See Dekel et al. (2007) for further discussion about the difference between independent and correlated interim rationalizability.

[6]To be precise, $\mathcal{S}$ maps the game induced by mechanism γ into a set of solutions. As the set of players, the type space, and utility functions are taken as given, for the sake of brevity let us just say $\mathcal{S}$ associates each mechanism γ with the set of its solutions $\mathcal{S}(\gamma)$ in the remainder of the paper.

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

For the remainder of the paper, we will refer to "full implementation" simply as "implementation" unless otherwise specified. We will moreover refer to the requirement that $\mu(\mathcal{S}(\gamma)) = f$ as the *uniqueness requirement*, as it demands all solutions of the mechanism yield the very same SCF.

We say a SCF $f \in F$ is *Bayesian Incentive Compatible* (BIC) whenever truthful reporting is an equilibrium in the direct mechanism (f, T). That is, whenever for all agents $i \in I$ and types $t_i, t_i' \in T_i$:

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

That is, when agent i of type t_i has no incentive to pretend to be of type t_i' in the direct mechanism associated with the SCF. We say f is BIC whenever it is BIC for all types of all agents.

Similarly, we say f is *Strict-if-Responsive Bayesian Incentive Compatible* (SIRBIC) for type t_i of agent $i \in I$ whenever it is BIC for $t_i \in T_i$ and the inequality above is strict for all $t_i' \neq t_i$ such that $f(t_i', t_{-i}) \neq f(t)$ for some $t_{-i} \in T_{-i}$. Again, we say f is SIRBIC whenever it is SIRBIC for all types of all agents.

We will derive most of our results about necessity of BIC by imposing a mild requirement on the solution concept $\mathcal{S}$. This requirement can be interpreted as requiring for each type t_i of agent i there exists a solution of the mechanism such that she has no incentive to play the strategy associated to a different type.

Definition 1 (Weak Response Consistency (WRC))**.** *We say a solution concept $\mathcal{S}$ satisfies WRC for a class of mechanisms $\tilde{\Gamma} \subseteq \Gamma$ whenever for all $\gamma \in \tilde{\Gamma}$ $i \in I$, $t_i \in T_i$ there exists $\sigma \in \mathcal{S}(\gamma)$ such that for all $t_i' \in T_i$:*

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

We say $\mathcal{S}$ satisfies WRC if it satisfies WRC for all $\gamma \in \Gamma$ such that $\mathcal{S}(\gamma) \neq \emptyset$.

It is worth noticing that this solution σ needs not be the same for all players i: as Kneeland (2022) highlights in Remark 2, this is because we do not require agents' expectations to be consistent anymore. As we allow expectations to be type-dependent, this solution need not be the same for any type t_i of player i either[8]. This fact highlights WRC is much weaker than incentive compatibility, which instead requires σ to be the same for all types

[8]About this point, see the discussion of weak IRM in Kunimoto et al. (2020).

of all players[9]. Moreover, notice Definition 1 directly implies that $\tilde{\mathcal{S}}$ is WRC for $\tilde{\Gamma} \subseteq \Gamma$ whenever there exist a WRC solution concept $\mathcal{S}$ such that $\mathcal{S}(\gamma) \subseteq \tilde{\mathcal{S}}(\gamma)$ for all $\gamma \in \tilde{\Gamma}$.

Finally, we 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 metrizable 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 SCF, mechanisms, and strategies be measurable functions.

2.1 A Sufficient Condition for WRC

Making use in a more explicit way of the definitions of E and R from Section 2 enables us to provide a sufficient condition on $\mathcal{S}$ for WRC that is both insightful and easy to check.

We say a solution concept $\mathcal{S}$ is *Response Consistent* (RC) for mechanism γ whenever $R \subseteq B$ and for all $i \in I$ and $t_i \in T_i$ there exists $e \in E(\gamma)$ and $\sigma \in R(e)$ such that $(\sigma_i, e_{i,t_i}) \in \mathcal{S}(\gamma)$[10]. It is immediate to see if $\mathcal{S}$ satisfies RC for mechanism γ, then it satisfies WRC for the same mechanism. We moreover say $\mathcal{S}$ is RC whenever it is RC for all $\gamma \in \Gamma$ such that $\mathcal{S}(\gamma) \neq \emptyset$.

Response Consistency is a rather mild requirement on $\mathcal{S}$ as it requires only that agents best respond to their expectations, and that these expectations about opponents (and agents' responses to them) could be justified as part of a solution to the mechanism. This means we can also interpret RC as demanding agents believe their opponents display a minimal level of rationality: type t_i of i responds and expects her opponent to respond to some expectation profile $e' \in E(\gamma)$, i.e. that $(\sigma_i, e_{i,t_i}) \in \mathcal{S}(\gamma)$.

To make the interpretation above clearer, it is instructive to consider an example in which RC does not hold. This is the case, for example, if we assume all players are either of level 1 or 0 in the models of de Clippel et al. (2019), Crawford (2021), and Kneeland (2022). Suppose that, for all level-0 agents, the anchor is to play a dominated strategy. Any level-1 agent i of type t_i knows then that the profile $(\sigma_i(t_i'), \alpha_{-i}(t_{-i}))$ will not be a solution to the mechanism, as it involves the play of a dominated action for all of i's opponents. These agents would then have no reason to pretend to be of type t_i' rather than t_i as they are aware this would not "fool" the planner into implementing a different outcome. Indeed,

[9]See also the discussion about Total Weak Response Consistency in Section 4.2.

[10]As $(\sigma_i, e_{i,t_i}) \in \mathcal{S}(\gamma)$ if and only if there exists $e' \in E(\gamma)$ such that $(\sigma_i, e_{i,t_i}) \in R(e')$, we can equivalently state RC as requiring that for all $i \in I$ and $t_i \in T_i$ there exists $e, e' \in E(\gamma)$ and $\sigma \in R(e)$ such that $(\sigma_i, e_{i,t_i}) \in R(e')$.

assuming all agents can be at least level 2 is crucial to ensure RC holds for all $\gamma \in \Gamma$ (Section 5.1).

3 A bilateral trading example

We can clarify the intuition behind BIC's necessity for full implementation of functions by considering 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, we summarize how level-k models of behavior work. Level-0 players of type t_i are naïve and (non-strategically) play some anchor $\alpha_i(t_i)$, which is exogenous to the model. Level-1 agents are instead 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 σ^1. For every level $k_i > 1$, agents of level k believe their opponents to be playing a *level-(k-1) consistent* strategy σ^{k_i-1}, and best respond accordingly. We say profile σ is a solution to a game γ whenever there exists a combination of levels $\{k_i\}_{i\in I}$ such that $k_i > 0$ for all $i \in I$ and σ_i is level-k_i consistent for all $i \in I$[11].

Suppose two risk-neutral parties trade an indivisible object with value c for the seller and v for the buyer, 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 a price $x = 0.5(a + b)$. Utility from not trading is 0 for both parties, while the utility from trading is $u_s = x - c$ and $u_b = v - x$ for the seller and the buyer, respectively.

In the discussion in Crawford (2021), we assume 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 $0.75 - \frac{1}{6}(2 + c)$ (as she believes her opponent is a level-0 agent), violating Bayesian Incentive Compatibility.

[11] As in de Clippel et al. (2019), each agent's type describes only her beliefs about the payoff-relevant state: as levels do not affect preferences, they are not part of the description of an agent's type.

However, the same function is not implementable if the two agents could both be of level $k = 2$. 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 a solution to the mechanism considered: the mechanism fails to implement f, however, as the two solutions lead to different outcomes. 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 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 μ has a solution σ^1 (so that σ_i^1 is a best reply to α_{-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 σ_i^2 to σ_{-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 σ^1 induces non-BIC f, there would then exist $i \in I$, $t_i, t_i' \in T_i$:

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

This implies σ^2 is not a best reply to σ^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 non-incentive compatible SCF[12].

This insight relies on the properties of level-k models, which are often solved recursively starting from the anchor. Quite surprisingly, this is not the case: in the following sections, we will prove the same holds for a much larger class of solution concepts. In particular, any model in which players best respond to their expectations and believe others are best responding to their own (possibly heterogeneous) expectations makes BIC necessary for

[12]It would still be possible for the mechanism to implement a social choice *set*: as a matter of fact, de Clippel et al. (2019) proves BIC is no longer necessary for level-k implementation in this case.

implementation. This class of solution concepts is larger than the classes of equilibrium or level-k reasoning ones, as each player and type may best respond to expectations different from the ones of their opponents and other types. Section 6.1 discusses how, for full implementation of functions, the impossibility result of Myerson and Satterthwaite (1983) generalizes to a large class of solution concepts, confirming its robustness even outside the rational expectations paradigm.

4 Results

We prove that BIC is still a necessary condition for implementation of functions if and only if the solution concept satisfies a novel property (Weak Response Consistency, Section 4.1). This property can be interpreted as requiring that for each type of each agent there exists a solution to the mechanism in which she does not want to imitate a different type. WRC is satisfied by several solution concepts that have been considered in the literature, with some notable exceptions (Section 5). It is not enough, however, to establish BIC is necessary for full implementation of sets (Section 4.2).

4.1 Full Implementation of Functions

There is a tight link between WRC and necessity of BIC for implementation of functions: BIC remains a necessary condition whenever the solution concept is WRC. Conversely, if f is BIC, $\mathcal{S}$ is WRC for the whole class of implementing mechanisms Γ^f.

Theorem 1. *If f is implementable in WRC $\mathcal{S}$, then it is BIC. If f is BIC and implementable in $\mathcal{S}$, then $\mathcal{S}$ is WRC for Γ^f.*

WRC solution concepts allow each agent to pretend to be of a different type: therefore, any implementable SCF must provide agents an incentive not to misreport their type. The result then follows as, due to the uniqueness requirement, the *same* SCF must incentivize all agents not to mimic a different type.

The full proof for the result is relegated to Appendix A. It is however instructive to discuss here a sketch of the argument for the if part, to appreciate how WRC and the uniqueness requirement of full implementation drive the final result. The key step of the proof involves noticing that whenever there exists a solution σ of mechanism γ such that

for type t_i and all $t_i' \in T_i$:

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

Then, as any f implemented by γ is such that $\mu(\sigma) = f$ by the uniqueness requirement of full implementability, the inequality above yields:

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

It is then immediate to notice WRC ensures that such a solution $\sigma \in \mathcal{S}$ exists for all $i \in I$ and $t_i \in T_i$ and, due to the uniqueness requirement, that all such solutions will yield f as an outcome. This is enough to establish that f is indeed BIC.

As argued below, the class of WRC solution concepts is rather broad, and it includes the level-k model of de Clippel et al. (2019), Bayes Nash Equilibrium, and Interim Correlated Rationalizability (Dekel et al., 2007). This list is not meant to be exhaustive of all WRC solution concepts: for other examples not considered in the literature so far see the discussion about Bayesian Correlated Equilibrium and Δ-rationalizability at the of Section 5.2[13].

4.1.1 Necessity of SIRBIC

It is also possible to use E and R in a more explicit way to prove that necessity of SIRBIC is a byproduct of the assumption that all best replies to an agent's expectations concur to form a solution to the mechanism, rather than the use of a non-equilibrium solution concept. This is the case, for example, in de Clippel et al. (2019) and Kunimoto et al. (2020).

Theorem 2. *Suppose f is BIC and implementable in $\mathcal{S}$. If $R = B$, then f is SIRBIC.*

That is, if all best replies to a profile of expectations are solutions to the mechanism (as is the case for the examples discussed in Section 5), then SIRBIC obtains for free from BIC and implementability.

4.2 Full Implementation of Sets

The results in Section 4.1 suggest that the necessity of BIC is robust even if we consider non-equilibrium solution concepts for the case of full implementation of functions. This section

[13]Although the tools described in this paper can be used to investigate the necessity of BIC for implementation in other solution concepts as well, such an endeavor falls beyond the scope of the present work.

considers implementation of social choice *sets* instead. de Clippel et al. (2019) and Kneeland (2022) already prove implementation of sets is more permissive than implementation of functions, as incentive compatibility of F is not necessary for implementation. Theorem 3 proves these positive results are due to the relaxation of the uniqueness requirement.

Theorem 3. *If F is implementable in WRC $\mathcal{S}$, then for all $i \in I$ and $t_i \in T_i$ there exists $f^{i,t_i} \in F$ that is BIC for i and t_i. Conversely, if F is implementable and there exists $f^{i,t_i} \in F$ that is BIC for i and t_i, then $\mathcal{S}$ is WRC for Γ^F.*

This result generalizes the standard Bayesian Incentive Compatibility constraint, showing that only a form of *partial* incentive compatibility is necessary for implementation of sets. Incentive constraints can be satisfied through a different function f^{i,t_i} for each agent and type[14]: a key implication is that the planner will be able to promise each type of each agent a different incentive f^{i,t_i} exploiting heterogeneity in expectations across agents and types. Conversely, BIC requires the same function f to satisfy the incentive constraints of all players and types, imposing $f^{i,t_i} = f^{j,t_j}$ for all $i, j \in I$, $t_i \in T_i$ and $t_j \in T_j$. This provides intuition as to why implementation of sets is much more permissive than the one of functions: as the planner is not restricted to a unique outcome for all solutions, she can decouple the incentives provided to each type of each player, possibly allowing for implementation of sets that do not contain any incentive compatible SCF.

We can also characterize the set of solution concepts that make BIC necessary for implementation when the uniqueness requirement is dropped. As the discussion of Theorem 3 suggests, this class will feature concepts in which beliefs are consistent across players and types.

Definition 2 (Total Weak Response Consistency (TWRC))**.** *We say a theory of expectations E is TWRC for a mechanism γ whenever for all $\sigma \in \mathcal{S}$, $i \in I$, $t_i, t_i' \in T_i$:*

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

TWRC requires each solution of the mechanism to be incentive compatible, and it is almost equivalent to BNE: the only difference is that for all $i \in I$, $\sigma_i(t_i)$ needs to be a better (rather than best) reply than $\sigma_i(t_i')$ to profile σ_{-i}. This comes very close to requiring rational expectations as well, as it entails all types of all agents have consistent expectations about the outcome that will prevail in the mechanism.

[14] A point similar to the one Kneeland (2022) makes about level-k models.

Theorem 4. *If F is implementable in TWRC $\mathcal{S}$, then it is BIC. If F is implementable and BIC, then $\mathcal{S}$ is TWRC for Γ^F.*

Therefore, necessity of BIC is a more "fragile" result whenever F is not a singleton, as it stems from very restrictive assumptions about the solution concept. This fragility is due to the fact a multi-valued F allows different players to believe different outcomes will prevail in the mechanism: therefore, it becomes no longer necessary for the *same* outcome to simultaneously provides incentives not to misrepresent their private information to each type and agent.

The fact that TWRC and WRC are equivalent for all mechanisms yielding the same outcome for all solutions confirms this intuition. It follows then that the BIC restriction on implementable social choice rules can be imputed in case of non-rational expectations more on the uniqueness requirement than on the solution concept $\mathcal{S}$.

Theorem 3 provides a weaker result than the one obtained from Kneeland (2022) for level-k implementation. In this case, we can make use of E in a more explicit way to bridge the gap between the two, by requiring expectations not to depend on each agent's type[15]: Section 5.1 checks this is indeed the case in Kneeland (2022)'s model.

Theorem 5. *If F is implementable in RC $\mathcal{S}$ and expectations are type-independent, for all $i \in I$ there exists $f \in F$ that is BIC for i.*

In other words, Theorem 5 is telling us that if expectations are constant with respect to i's type, then the same SCF f must provide all types $t_i \in T_i$ with an incentive not to mimic another type[16]. As for the difference between Theorem 1 and Theorem 3, the comparison of Theorem 3 and Theorem 5 highlights how heterogeneity in expectations leads to a larger class of implementable social choice rules.

5 Examples

As argued above, RC does not seem to be a particularly restrictive condition. It is indeed satisfied by various solution concepts proposed in the literature: BNE (Jackson, 1991),

[15]It would be possible to express the results above in terms of a modified WRC condition as well, by substituting the quantifier "for all $i \in I$ and $t_i \in T_i$" with "for all $i \in I$". However, stating the results in terms of expectations seems to be more intuitive.

[16]Notice that for the result to go through, the argument in the proof of Theorem 5 requires only that there exists one type-independent expectation in $E(\gamma)$.

level-k reasoning (de Clippel et al., 2019; Kneeland, 2022), and Interim Correlated Rationalizability (Kunimoto et al., 2020). It is also satisfied for a weaker solution concept than BNE, which requires agents to share the same (not necessarily correct) expectations about their opponents.

WRC is not satisfied, instead, by the cursed equilibrium model of Eyster and Rabin (2005) and by the model of projection bias of Gagnon-Bartsch et al. (2021). This is due to the fact that, in these models, the profile of strategies played by each agent as a response to their expectations and their expectations themselves do not are generally not a solution to the mechanism. These novel examples, together with the level-k model with no level 2 agents of Crawford (2021), serve to confirm that WRC does indeed have bite.

5.1 Level-k model

The discussion in this section builds on the models of de Clippel et al. (2019) and Kneeland (2022), where it is assumed combinations of agents' levels k_i are possible, up to an upper bound $\bar{k} \geq 2$. That is, $k_i \leq \bar{k}$ for all $i \in I$.

Let $\alpha : \Gamma \to \Sigma$ be any correspondence assigning a profile of anchors to each mechanism $\gamma \in \Gamma$. We will denote the set of best replies to these anchors as the set of *level-1 consistent* strategies $S_i^1(\gamma|\alpha)$. Similarly, we will denote the set of best replies to *level-(k_i-1) consistent* strategy profiles as the set of *level-k_i consistent* strategies $S_i^{k_i}(\gamma|\alpha)$.

We can then characterize the set of solutions of each mechanism γ by setting $R = B$ and $E = E^{K,\alpha}$, where:

$$E^{K,\alpha}(\gamma) = \{e \in \mathcal{E} : e_{i,t_i} \in \{\alpha_{-i}(\gamma)\} \cup \{\cup_{1 \leq k_i \leq \bar{k}} S_{-i}^{k_i - 1}(\gamma|\alpha)\}, e_{i,t_i} = e_{i,t_i'}, \text{ for all } i \in I, t_i, t_i' \in T_i\}$$

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 $k_i - 1$ ($e_i \in \cup_{1 \leq k_i \leq K} S_{-i}^{k_i - 1}(\mu|\alpha)$). It is immediate to notice any strategy profile such that each player's strategy is level-k_i consistent for $k_i \geq 1$ is a solution of the mechanism[17].

This solution concept satisfies RC for all $\gamma \in \Gamma$. As $\bar{k} \geq 2$, $E^{K,\alpha}(\gamma)$ contains at least one e such that $e_i \in S_{-i}^1(\gamma|\alpha)$. Consider then, for all $i \in I$ and $t_i \in T_i$, any $\sigma \in B(e) = R(e)$. It is then clear that $(\sigma_i, e_i) \in \mathcal{S}(\gamma)$ as σ_i is a level-2 consistent strategy and e_i is a profile of level-1 consistent strategies.

[17] It is also worth mentioning in this case expectations are type-independent, allowing us to derive slightly stronger result for full implementation of social choice sets (Theorem 5).

The assumption $\bar{k} \geq 2$ (de Clippel et al., 2019; Kneeland, 2022) is then useful to exclude pathological cases in which a player expects her opponents just to play their anchor. This seems to account for the differences between the aforementioned papers and Crawford (2021), which proves it is possible to implement non-incentive compatible rules. This possibility result stems from the fact Crawford (2021) considers a setup with no level-2 players, allowing for the possibility RC may not hold.

5.2 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 SCF.

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

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

As argued in the paper, by Tarski's theorem, there exists a largest fixed point of b which is denoted as $C^{\gamma(T)}$. 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:

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

$$R(e) = \{\sigma \in \Sigma : \sigma \in B(e), |\text{supp}(\sigma)| = 1\}$$

This follows from the fact $\mathcal{S}(\gamma) = C^{\gamma(T)}$ for all $\gamma \in \Gamma$. As a matter of fact, $\sigma \in B(e)$ for $e \in E(\gamma)$ implies the unique profile s in σ's support is a rationalizable profile of actions, and thus that $\sigma \in C^{\gamma(T)}$ and $\mathcal{S} \subseteq C^{\gamma(T)}$. Conversely, if $\sigma \in C^{\gamma(T)}$, for all $i \in I$ and $t_i \in T_i$ there exists a belief $\lambda \in \Delta(T_{-i} \times S_{-i})$ to which $\sigma_i(t_i)$ is a best reply: setting $e_{i,t_i} = \lambda$ is then enough to achieve $C^{\gamma(T)} \subseteq \mathcal{S}$, concluding the argument.

It can also be shown that Interim Correlated Rationalizability satisfies RC for a large class of mechanisms $\gamma \in \Gamma$ as well: in particular, the ones in which $B_{i,t_i'}(e_{i,t_i}) \neq \emptyset$ for all

$i \in I$ and $t_i, t_i' \in T_i$. Consider any solution $\sigma \in \Sigma$ and suppose $B_{i,t_i'}(\sigma_{-i}) \neq \emptyset$ for all $i \in I$ and $t_i' \in T_i$. Then, for each $\tilde{\sigma}_i$ such that $\tilde{\sigma}_i(t_i) \in B_{i,t_i}(\sigma_{-i})$, $(\tilde{\sigma}_i, \sigma_{-i}) \in \mathcal{S}(\gamma)$: this follows from the fact $\sigma \in \mathcal{S}(\gamma)$ entails σ_{-i} is rationalizable for all agents $j \neq i$ and $\tilde{\sigma}_i$ is rationalized by the belief i's opponents are going to play σ_{-i}.

There are two important cases in which this assumption holds: if either the mechanism or the set of alternatives A are finite, and if $\mathcal{S}(\gamma)$ is convex[18]. To see why convexity matters, fix arbitrary $i \in I$ and $t_i \in T_i$ and consider any $\sigma \in \mathcal{S}(\gamma)$ and $e \in E(\gamma)$ such that $\sigma \in B(e)$. For any $t_{-i} \in T_{-i}$ and $\sigma'_{-i}(t_{-i})$ in the support of e_{i,t_i}, $(\sigma_i(t_i), \sigma'_{-i}(t_{-i}))$ is a rationalizable profile of actions. This follows from the fact $\sigma_i(t_i)$ is a best reply to e_{i,t_i} for type t_i and $\sigma'_{-i}(t_{-i})$ belongs to the support of e_{i,t_i}. As the set of solutions is convex, then $(\sigma_i(t_i), e_{i,t_i}(t_{-i}))$ is rationalizable in state t as well. As our choice of t was arbitrary, $(\sigma_i, e_{i,t_i}) \in \mathcal{S}(\gamma)$ and thus $\mathcal{S}(\gamma)$ is RC.

We can also notice that a similar argument applies 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 Δ-rationalizability (Battigalli and Siniscalchi, 2003).

5.3 Bayesian Nash equilibrium and refinements

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

$$E^{BN}(\gamma) = \{e \in \mathcal{E}(\gamma) : \exists\ \sigma \in \times_{i \in I}\Sigma_i \text{ s.t. } e_{i,t_i} = \sigma_{-i} \text{ for all } i \in I, t_i \in T_i, \sigma \in B((\sigma_{-i})_{i \in I})\}$$

$$R^{BN}(e) = \{\sigma \in B(e) : \sigma_{-i} = e_i \text{ for all } i \in I\}$$

It is then clear the set of BNE is equal to $R^{BN}(E^{BN}(\gamma)) = \mathcal{S}^B N(\gamma)$. As a matter of fact, if $\sigma \in \mathcal{S}^B N(\gamma)$ we have that $\sigma_i \in B(\sigma_{-i})$ for all $i \in I$. On the other hand, if σ is a BNE it is immediate to notice that $(\sigma_{-i})_{i \in I} \in E^{BN}(\gamma)$ and thus that $\sigma \in R^{BN}(e)$. Moreover, as long as $E^{BN}(\gamma) \neq \emptyset$, $R^{BN}(E^{BN}(\gamma)) \neq \emptyset$ as well as R^{BN} just selects, among all profiles of best responses, the one satisfying rational expectations. $\mathcal{S}^{BN}$ also satisfies RC for all $\gamma \in \Gamma$ as for all $i \in I$ and $t_i \in T_i$, expectation profile $e' = (\sigma_{-i})_{i \in I}$ is such that $(\sigma_i, e_{i,t_i}) = \sigma \in B(e')$. The same argument applies to refinements of BNE as well (as undominated BNE), as they all satisfy the rational expectations assumption.

[18] This is the case, for example, for all the mechanisms implementing a SCF f.

5.4 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 theory of expectations, dropping this last correctness requirement. Let $R^{SE} = B$ and $E^{SE}(\gamma) = E^{BN}(\gamma)$. It is clear this solution concept satisfies RC, as $\mathcal{S}^{BNE}(\gamma) \subseteq \mathcal{S}^{SE}(\gamma)$ for all $\gamma \in \Gamma$. It is then enough to follow the steps in the previous subsection to establish the result.

5.5 Bayes Correlated Equilibrium

Bergemann and Morris (2016) propose Bayesian Correlated Equilibrium (BCE) as an extension to incomplete information of correlated equilibrium (Aumann, 1974). The set of BCE of the game induced by mechanism γ can be captured by setting $\mathcal{S}^{BCE} = (E^{BCE}, R^{BCE})$ with[19]:

$$E^{BCE}(\gamma) = \{e \in \mathcal{E}(\gamma) : \exists\ \sigma \in \Sigma \text{ s.t. } e_i = \sigma_{-i} \text{ for all } i \in I, \sigma \in B((\sigma_{-i})_{i \in I})\}$$

$$R^{BCE}(e) = \{\sigma \in B(e) : e_i = \sigma_{-i} \text{ for all } i \in I\}$$

Let $\sigma_{-i} : T \to \Delta(S_{-i})$ denote the mixed action played by agents other than i. We can then notice the definition of E^{BNE} and R^{BNE} entail $\sigma \in \mathcal{S}^{BCE}$ if and only if for all $i \in I$, $t_i \in T_i$ and $s_i' \in \Delta(S_i)$:

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

So that $\sigma \in \mathcal{S}^{BCE}$ if and only if σ is a BCE as well.

As $\mathcal{S}^{BNE} \subseteq \mathcal{S}^{BCE}$, it follows immediately that $\mathcal{S}^{BCE}$ is WRC. It follows then from Theorem 1 that any SCF f which is fully implementable in $\mathcal{S}^{BNE}$ will be BIC, and from Theorem 3 that any SCS F which is fully implementable in $\mathcal{S}^{BNE}$ will be partially BIC.

5.6 Cursed equilibrium

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

$$E^{CE}(\gamma) = \{e \in \mathcal{E}(\gamma) : \exists\ \sigma \in \times_{i \in I}\Sigma_i \text{ s.t. } e_{i,t_i} = \sigma_{-i} \text{ for all } i \in I, t_i \in T_i, \sigma \in B((\bar{\sigma}_{-i})_{i \in I})\}$$

[19]The

$$R^{CE}(e) = \{\sigma \in B(e) : \bar{\sigma}_{-i} = e_i\}$$

Where:

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

It is possible to prove this solution concept is WRC for all mechanisms γ if agents have private values, but it is not otherwise. The intuition for this result is that the distribution of actions agents expect from their opponents is not, generally speaking, rationalizable as a best reply to their own expectations.

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

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

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

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

Which, as σ is a solution, concludes the proof.

However, the same is not valid for all mechanisms γ if an agent's payoff depends on the type of her opponents. Intuitively, this stems from the fact the profile $(\sigma_i, \bar{\sigma}_{-i})$ is typically not a solution to the mechanism, violating RC (and WRC, as argued below). For example, for any $\chi \in (0, 1]$, 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 $\zeta \in (2, \frac{2}{1-\chi})$[20]. The only cursed equilibrium of this game is for type 1 to play A and for type -1 to play B. To prove it, consider any solution σ of the game. Then:

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

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

The payoff of B is always 0 for either player, while the payoff from playing A is:

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

Type 1 will play A with probability 1 as long as:

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

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

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

There is therefore a pure cursed equilibrium in which both agents play A if their type is $t_i = 1$ and B otherwise. Moreover, this is the unique cursed equilibrium of the game and it does not satisfy WRC. 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 to mimic type $t_i = -1$ if she were not ignoring the correlation between her opponents' strategies and types: cursed equilibrium then allows for the implementation of non-BIC social choice functions.

5.7 Naïve Bayesian Equilibrium

Gagnon-Bartsch et al. (2021) propose a model of taste projection, i.e. the tendency of agent to believe their opponents' valuations for an object are similar to their own. The associated solution concept, Naïve Bayesian Equilibrium (NBE), is meant to capture the idea that agents play according to the BNE of a *perceived* game in which beliefs are distorted by taste projection.

In order to keep the discussion simple, let agents' value be private and independent[21]. Let then $BNE(\gamma, \hat{\tau}(t_i))$ denote the set of pure-strategy BNEs of mechanism γ when agents believe their opponents' values are determined according to the random variable $\hat{\tau}(t_i) = \chi t_i + (1-\chi)\tau$ rather than the true random variable τ. We can then define $\mathcal{S}^{NBE} = (E^{NBE}, R^{NBE})$ as follows:

$$E^{NBE}(\gamma) = \{e \in \mathcal{E}(\gamma) : e_{i,t_i} = \sigma_{-i} \text{ for all } i \in I, t_i \in T_i, \text{ where } \sigma \in BNE(\gamma, t_i)\}$$

$$R^{NBE}(e) = \{\sigma \in \Sigma : (\sigma_i, e_i) \in BNE(\gamma, \cdot) \text{ for all } i \in I\}$$

[21] More general versions of this model can be accommodated in the framework presented in Section B.1.

Let us say (A, T, I) is an *economic environment* whenever for all $t \in T$, $i, j \in I$ and $a \in A$, there exist $b, c \in A$ such that $u_i(b, t) > u_i(a, t)$ and $u_j(c, t) > u_j(a, t)$. Say also (A, T, I) satisfies *single crossing* whenever $u_i(a, t) \geq u_i(b, t)$ and $u_i(a, t') \geq u_i(b, t')$ for all $i \in I$ imply there exists $c \in A$ such that $u_i(a, t) \geq u_i(c, t)$ and $u_i(c, t) > u_i(a, t)$.

Theorem 6 shows Naïve Bayesian Equilibrium is generally not WRC.

Theorem 6. *Let $|I| \geq 3$. If $\chi = 1$ and (A, T, I) satisfies the single crossing and economic environment assumptions, any SCF f can be implemented in NBE via the Maskin mechanism (Maskin, 1999).*

For $\chi = 1$, the perceived mechanism is a complete information game. Then, if (A, I, T) satisfies the economic environment and single crossing assumptions, any SCF f is implementable via the canonical mechanism in Maskin (1999) as no-veto power and Maskin-motonocity are always trivially satisfied. Moreover, the NBE will be unique, as all BNEs of the Maskin mechanism prescribe the same action for agent i of type t_i. Under these restrictions on the preference domain, all SCF are then implementable whether they are BIC or not. By Theorem 1 this entails that, in general, NBE is not a WRC solution concept.

6 Applications

The results in previous sections allow us to extend results stemming from necessity of BIC for implementation to all WRC solution concepts. We take as examples three classical results from the mechanism design literature: impossibility of efficient bilateral trade (Myerson and Satterthwaite, 1983), impossibility of full surplus extraction in auctions and the Revenue Equivalence Theorem (Myerson, 1981). Our results confirm the economic intuition behind these results extends to a wide range of boundedly rational setups.

6.1 Myerson-Satterthwaite's Impossibility Theorem

Myerson and Satterthwaite (1983) show that efficient bilateral trade may be impossible in presence of private information, unless the planner steps in to cover some of the losses the agents face. As this results relies on necessity of BIC for implementation in BNE, Theorem 1 allows us to extend it to all WRC solution concepts.

As in Myerson and Satterthwaite (1983), we consider a bargaining problem in which two agents (a buyer B and a seller S) bargain over the sale of an indivisible object which each agent values at t_i distributed according to $p_i : T_i \to [0, 1]$. We assume p_i admits a

continuous and positive pdf over the interval $[a_i, b_i]$, with $(a_S, b_S) \cap (a_B, b_B) \neq \emptyset$. We also assume t_B is independent of t_S, that each agent knows her valuation and how the valuation of the other agent is distributed. The set of alternatives consists of all pairs (q, x), where $q \in [0, 1]$ represents the probability that trade will happen, and x indicates the amount transferred from the buyer to the seller. Bernoulli utilities $u_i : A \times T_i \to \mathbb{R}$ are additively separable in money and the value of the object, and agents are risk neutral.

Under these assumptions, Myerson and Satterthwaite (1983) prove an implementing mechanism that assigns an object to the agent who values it the most is unable to ensure voluntary participation from both agents. Formally, we say a SCF is *ex-post efficient* if it allocates the object with probability 1 to the agent who values it the most, i.e. $q(t) = 1$ whenever $t_B > t_S$ and $q(t) = 0$ whenever $t_B < t_S$. We say instead f is *individually rational* whenever $u_i(f(t), t) = q(t)t_i - x(t) \geq 0$ for all $i \in I$ and $t \in T$.

The proof of Myerson and Satterthwaite (1983) relies on showing there exists no SCF f that is simultaneously individually rational, ex-post efficient, and BIC. It then follows from Theorem 1 that:

Corollary 1. *If f is individually rational and ex-post efficient, it is not fully implementable in any RC $\mathcal{S}$.*

Myerson and Satterthwaite (1983) highlight it is impossible to find an ex-post efficient and individually rational SCF that is also incentive compatible for all types and agents *at the same time*. This result extends the negative ones de Clippel et al. (2019) and Crawford (2021) obtain for full implementation of social choice function in level-k reasoning.

Kneeland (2022) shows instead it is possible to fully implement an efficient and individually rational social choice *set*. Each agent can believe a different solution of the mechanism will obtain when F is not a singleton, allowing the planner to decouple the incentives she provides. That is, F must contain one SCF that is incentive compatible for each agent and type but needs not to contain a SCF that is incentive compatible for all types of all agents at the same time.

6.2 Impossibility of Full Surplus Extraction

If $\mathcal{S}$ is WRC, the planner cannot implement an auction extracting all expected surplus from agents unless she excludes lower-ranked types from winning the object.

Suppose the planner is tasked with designing a mechanism to allocate a single unit of an indivisible object in exchange for the payment of a fee. Let the set of alternatives be

defined as:

$$A = \{(q, x) \in [0,1]^I \times \mathbb{R}^I : \sum_{i \in I} q_i \leq 1\}$$

That is, $f(t)$ assigns to each agent some probability to win the object and a (non-contingent) monetary transfer. For a given f, denote as $q_i^f(t)$ the probability agent i receives the object and $x_i^f(t)$ the associated transfer to the planner for the agent getting the object. Assume moreover that $T \subseteq \mathbb{R}^I$, and types are determined by a commonly known the joint distribution $p : T \to [0,1]$. The value of the object for i is determined according to a function v_i that is strictly increasing in i's type, and Bernoulli utilities takes the additively separable form $u_i(t) = v_i(t) - x_i$.

We then say a SCS F is *fully extractive* whenever $x_i^f(t) = q_i^f(t)v_i(t)$ for all $t \in T$ and $f \in F$. Moreover, we say F is *inclusive* whenever for all $f \in F$, $i \in I$ there exists $t_i' \in T_i$ and $t \in T$ such that $t_i' > t_i$ and $q_i^f(t) > 0$. In other words, inclusivity requires that f does not prevent all types t_i that are ranked lower than t_i' from getting the object with positive probability for all type profiles t_{-i} of other agents. This is the case, for example, for ex-post efficient allocation rules.

We can then prove there exists a tradeoff between inclusivity and total surplus extraction.

Corollary 2. *If F is fully extractive and inclusive, then it is not implementable in any WRC $\mathcal{S}$.*

The result follows from the fact that inclusivity and complete extraction of surplus entail each type has an incentive to pretend the object is worth less for her than it actually is. This creates a tension with implementability in a WRC solution concept, which implies instead there exists at least one SCF in F providing each agent with the incentive not to misrepresent her type. This should be contrasted with the result in the previous example, which follows instead from the fact that the same SCF has to be simultaneously incentive compatible for all types of all agents as in the application above. The impossibility faced in this application is therefore harder to escape than the one of Myerson and Satterthwaite (1983).

6.3 A Revenue Equivalence Theorem

Our results also allow us to extend Myerson's (1981) fundamental result about revenue equivalence of different auction formats to all SCF that are fully implementable in a WRC solution concept.

As in Myerson (1981), let us assume agents' values are drawn from set $[a_i, b_i] \subseteq \mathbb{R}_0^+$ according to some commonly known cdf p, that agents are risk neutral, that their utility is additively separable in money and the value of the object and that $v_i(t)$ is differentiable in t_i for all $i \in I$ (this is the case, for example, if $v_i(t) = t_i$).

Let $\bar{q}_i^f(t_i)$, $\bar{v}_i(t_i)$ and $\bar{x}_i^f(t_i)$ denote respectively the average probability of winning, value of the object and transfer for an agent of type t_i. We say a SCF $f = (q, x)$ is *differentiable* if both $\bar{q}_i$ and $\bar{x}_i$ are differentiable in t_i for all $i \in I$ almost everywhere, and that two SCF f and $\tilde{f}$ are *assignment-equivalent* if $q^f = q^{\tilde{f}}$ almost everywhere. Notice that f is differentiable whenever we assume it is ex-post efficient and that agents' values are independently and identically distributed according to a cdf p, as in that case $\bar{q}(t_i) = p^{n-1}(t_i)$.

Corollary 3. *If differentiable and assignment-equivalent SCF f and $\tilde{f}$ are fully implementable in WRC $\mathcal{S}$, then $\bar{x}_i^f(t_i) - \bar{x}_i^f(a_i) = \bar{x}_i^{\tilde{f}}(t_i) - \bar{x}_i^{\tilde{f}}(a_i)$ for all $i \in I$.*

Corollary 3 establishes a generalized version of the standard Revenue Equivalence Theorem of Myerson (1981), stating the revenue of a given SCF f is determined by its allocation probability q up to an additive constant $\bar{x}^f(a_i)$. If we standardize the average payment of type a_i to 0, we obtain the familiar result that any two rules f and $\tilde{f}$ that are fully implementable in RC $\mathcal{S}$ (and their associated implementing mechanisms: e.g., auctions) will yield the same ex-ante revenue to the planner unless they differ in the probability with which each type gets allocated the object. This fact entails, for example, that all ex-post efficient SCF must yield the same revenue to the planner when p is atomless.

References

Artemov, G., T. Kunimoto, and R. Serrano (2013). Robust virtual implementation: Toward a reinterpretation of the wilson doctrine. *Journal of Economic Theory 148*(2), 424–447.

Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics 1*(1), 67–96.

Barlo, M. and N. A. Dalkıran (2022). Behavioral implementation under incomplete information. Working paper.

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

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

Bergemann, D. and S. Morris (2016). Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics 11*(2), 487–522.

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

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

de Clippel, G., R. Saran, and R. Serrano (2019, 06). Level-k Mechanism Design. *The Review of Economic Studies 86*(3), 1207–1227.

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

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

Gagnon-Bartsch, T., M. Pagnozzi, and A. Rosato (2021, October). Projection of private values in auctions. *American Economic Review 111*(10), 3256–98.

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

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–291.

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

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

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York.

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

McKelvey, R. D. and T. R. Palfrey (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior 10*(1), 6–38.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research 6*(1), 58–73.

Myerson, R. B. and M. A. Satterthwaite (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory 29*(2), 265–281.

Oury, M. and O. Tercieux (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.

Radner, R. (1980). Collusive behavior in noncooperative epsilon-equilibria of oligopolies with long but finite lives. *Journal of Economic Theory 22*(2), 136–154.

Saran, R. (2011). Menu-dependent preferences and revelation principle. *Journal of Economic Theory 146*(4), 1712–1720.

Sen, A. K. (1971). Choice functions and revealed preference. *Review of Economic Studies 38*(3), 307–317.

Simon, H. A. (1955, 02). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics 69*(1), 99–118.

Zambrano, E. (2008). Epistemic conditions for rationalizability. *Games and Economic Behavior 63*(1), 395–405.

Appendix A Proofs

Proof of Theorem 1. Suppose f is implementable in $\mathcal{S}$ via mechanism $\gamma = (\mu, S)$, and that $\mathcal{S}$ is WRC for γ. Then $\mathcal{S}(\gamma) \neq \emptyset$ and there exists $\sigma \in \mathcal{S}(\gamma)$ such that:

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

As $\sigma \in \mathcal{S}(\gamma)$, implementability of f yields $\mu(\sigma) = f$. Therefore, for $i \in I$ and $t_i, t'_i \in T_i$:

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

As our choice of i, t_i and t'_i was arbitrary, this is enough to establish f is BIC.

Conversely, suppose f is BIC and implementable in $\mathcal{S}$ via mechanism $\gamma = (\mu, S)$. We have then that for all $t'_i \in T_i$ and $i \in I$:

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

By implementability, there exists $\sigma \in \mathcal{S}(\gamma)$ such that $\mu(\sigma) = f$ and thus:

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

Concluding the proof. □

Proof of Theorem 2. To prove we can strengthen it to SIRBIC, proceed by contradiction and suppose that, indeed, the incentive constraint in the proof of Theorem 1 holds with equality. Define then $\tau : T \to \Sigma$ as agreeing with σ except for the fact $\tau(t) = \sigma(t_i', t_{-i})$ for all $t_{-i} \in T_{-i}$. As σ is a solution to the mechanism, there exist e such that $\sigma \in B(e)$. As τ yields the same expected utility as σ conditional on expectations e, $\sigma \in B(e)$ implies $\tau \in B(e)$. Then by the definition of implementation above, we have for all $t_{-i} \in T_{-i}$:

$$f(t_i, t_{-i}) = \mu(\tau(t)) = \mu(\sigma(t_i', t_{-i})) = f(t_i', t_{-i})$$

This concludes the proof. □

Proof of Theorem 6. Let us first derive the set of BNE of the Maskin mechanism (μ, S) associated with f under the perceived distribution of types. Let $S = \times_{i \in I} S_i$, with $S_i = (T, A, \mathbb{N})$. The outcome function μ is as follows:

- **Rule 1:** if $s_i = (t, f(t), 0)$ for all $i \in I$ and $\mu(s) = f(t)$
- **Rule 2:** if $s_j = (t', a, \mathbb{N})$ and $s_i = (t, f(t), 0)$ for all $i \neq j$, then $\mu(s) = a$ if $u_i(f(t), t) \geq a$ and $\mu(s) = f(t)$ otherwise
- **Rule 3:** $\mu(s) = a$ otherwise, where a is the outcome reported by the agent with the lowest index among those that reported the highest integer

As (A, T, I) satisfies the economic environment assumption, the same argument as in Maskin (1999) implies that $\sigma(t)$ does not falls under Rule 2 or Rule 3 for all $t \in T$: if that was the case, at least one agent could report a higher integer than their opponents and achieve $a \in A$ such that $u_i(a, t) > u_i(\mu(\sigma(t)), t)$. Therefore, for all states $t \in T$, $\sigma(t)$ falls under Rule 1.

Moreover, in any BNE σ, $\sigma_i(t_i) = (t, f(t), 0)$ for all $i \in I$ and $t_i \in T_i$. As σ falls under Rule 1, then there exists $t' \in T$ such that $\sigma_i(t_i) = (t', f(t'), 0)$ for all $t_i \in T_i$ and $i \in I$. For the sake of contradiction, suppose now that $t' \neq t$. As σ is an equilibrium, it must hold for all $i \in I$ that $u_i(f(t'), t) \geq u_i(a, t)$ for all $a \in A$ such that $u_i(f(t'), t') \geq u_i(a, t')$: otherwise, at least one agent i could play $s_i' = (t', a, 0)$ and obtain $u_i(a, t) > u_i(f(t'), t)$. By single crossing,

for all $i \in I$ there exists $z \in A$ such that $u_i(f(t'), t') \geq u_i(z, t')$ and $u_i(z, t) > u_i(f(t'), t)$. It would then be profitable for type i to deviate to $s'_i = (t, z, 0)$, contradicting the premise that σ is an equilibrium strategy. It remains to show no type t_i of each $i \in I$ has a profitable deviation from $\sigma(t)$: as i can induce a only if $u_i(\mu(\sigma(t)), t) \geq u_i(a, t)$, this concludes the proof. □

Proof of Theorem 3. Suppose F is implementable in WRC $\mathcal{S}$ via mechanism $\gamma = (\mu, S)$ with $\mathcal{S}(\gamma) \neq \emptyset$. Then for each $i \in I$ and $t_i \in T_i$ there exists $\sigma \in \mathcal{S}(\gamma)$ such that for all $t'_i \in T_i$:

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

As $\sigma \in \mathcal{S}(\gamma)$, implementability of F yields $\mu(\sigma) = f$ for some $f \in F$. Therefore, for all $t'_i \in T_i$:

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

Which is enough to prove f is BIC for agent i and type t_i.

As for the converse, suppose F is implementable in $\mathcal{S}$ via mechanism γ and that for all $i \in I$ and $t_i \in T_i$ there exists $f \in F$ that is BIC for agent i and type t_i. Then for each such f, i and t_i there exists a solution σ such that $f = \mu(\sigma)$: a simple substitution in the BIC inequality yields WRC holds. □

Proof of Theorem 4. If F is implementable in $\mathcal{S}$, then any $f \in F$ is such that $f = \mu(\sigma)$ for $\sigma \in \mathcal{S}$. As $\mathcal{S}$ is TWRC, it is immediate to establish f is BIC from the definition of TWRC by substituting $f = \mu(\sigma)$. Conversely, suppose F is implementable in $\mathcal{S}$. As any $\sigma \in \mathcal{S}$ is such that $\mu(\sigma) \in F$, TWRC follows immediately from the fact all functions $f \in F$ are BIC. □

Proof of Theorem 5. Suppose F is implementable in RC $\mathcal{S}$, and that E is type-independent. By RC, for all $i \in I$ there exists $e \in E(\gamma)$ and $\sigma \in R(e)$ such that $(\sigma_i, e_i) \in \mathcal{S}(\gamma)$. Then for all $t'_i \in T_i$ we have by $\sigma \in R(e) \subseteq B(e)$:

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

For each $i \in I$, let $f = \mu \circ (\sigma_i, e_i)$. By RC, $\mu \circ (\sigma_i, e_i) \in F$, so $f \in F$. Moreover:

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

Entailing f is BIC for all types of agent i. This concludes the proof. □

Proof of Corollary 2. We will show that supposing F is implementable in WRC $\mathcal{S}$ leads to a contradiction. Consider any agent $i \in I$. By inclusivity, there exists types $t_i, t_i' \in T_i$ such that $q_i^f(t_i', t_{-i}) > 0$ and $t_i > t_i'$. By WRC and Theorem 3, we then know if F is implementable in $\mathcal{S}$, for all $i \in I$ and $t_i \in T_i$ there exists $f \in F$ that is BIC for i and t_i. Therefore, for all $i \in I$, $t_i \in T_i$ and $t_i' < t_i$:

$$0 = \int_{T_{-i}} (q_i^f(t)v_i(t) - q_i^f(t)v_i(t))\, dp_i(t_{-i}|t_i) \geq \int_{T_{-i}} q_i^f(t_i', t_{-i})(v_i(t) - v_i(t_i', t_{-i}))\, dp_i(t_{-i}|t_i)$$

As v_i is strictly increasing in i's type and F is inclusive:

$$\int_{T_{-i}} q_i^f(t_i', t_{-i})(v_i(t) - v_i(t_i', t_{-i}))\, dp_i(t_{-i}|t_i) > 0$$

This inequality contradicts the fact f is BIC for i and t_i, concluding the proof. □

Proof of Corollary 3. As f is implementable in $\mathcal{S}$ WRC, it is BIC. Then, $t_i^* = t_i$ must maximize the payoff function $\bar{q}^f(t_i^*)\bar{v}_i(t_i) - \bar{x}_i^f(t_i^*)$. A necessary condition for a maximum is that the first derivative with respect to v_i of this function is null at t_i, i.e. that $\frac{\partial \bar{x}^f(t_i)}{\partial t_i} = \frac{\partial \bar{q}^f(t_i)}{\partial t_i}\bar{v}_i(t_i)$. By the fundamental theorem of calculus:

$$\bar{x}(t_i) - \bar{x}(a_i) = \int_{a_i}^{t_i} \bar{v}_i(t_i') \frac{\partial \bar{q}^f(t_i')}{\partial t_i}\, dt_i'$$

An analogous reasoning for $\tilde{f}$ and assignment-equivalence yield $\bar{x}^f(t_i) - \bar{x}^f(a_i) = \bar{x}^{\tilde{f}}(t_i) - \bar{x}^{\tilde{f}}(a_i)$, concluding the proof. □

Appendix B Extensions of the Model

B.1 Non-standard Choice Functions

This section relaxes the assumption agents best respond to their expectations, generalizing the results above beyond the domain of von Neumann–Morgenstern preferences.

We can interpret the revelation principle as saying some lotteries in the choice set of an agent in the direct mechanism alternatives but not in the direct one can be safely neglected as they are "not relevant". Formally, this requires that restricting the choice set of an agent of type t_i to the set of lotteries that would be a solution to the mechanism for some type $t_i' \in T_i$ does not affect their choice. This will require us to impose some form of Contraction Consistency, or Independence of Irrelevant Alternatives (see, for example, Property α of Sen (1971)). In the argument below, we will only maintain the assumption that agents are consequentialist, i.e. their choices depend only on the set of alternatives they choose from[22].

As in Saran (2011) and Barlo and Dalkıran (2022), we model individual strategic decisions as choices over a set of *interim Anscombe-Aumann acts* (IAA acts) $x_i : T_{-i} \to \Delta(A)$. Denoting as $\mathcal{A}$ the set of all possible acts, we can then define a choice function C_{i,t_i} as a mapping $C_{i,t_i} : 2^{\mathcal{A}} \to 2^{\mathcal{A}}$ such that $C_{i,t_i}(X) \subseteq X$ for all $X \subseteq \mathcal{A}$. As in Barlo and Dalkıran (2022) and differently from Saran (2011), we do not assume C_{i,t_i} is generated by a menu-dependent preference order.

Notice that for any given $s_i \in \Delta(S_i)$ and $e \in E(\gamma)$, the function $\mu(s_i, e_{i,t_i})$ is an IAA act. We can then denote the set of acts agent i of type t_i chooses from (or believes to be choosing from) given their expectations as:

$$O_i(\sigma_{-i}) = \{x_i \in \mathcal{A} : x_i = \mu(s_i, \sigma_{-i}), s_i \in \Delta(S_i)\}$$

As in the previous sections, we say $s_i \in \Delta(S_i)$ is a reply for type t_i' to expectations e_{i,t_i} whenever $s_i \in R_{i,t_i'}(e) \subseteq \mu^{-1}(C_{i,t_i'}(O_i(e_{i,t_i})))$. That is, the outcome of the strategies chosen as a response to e_{i,t_i} is a subset of the choices prescribed by $C_{i,t_i'}$ on the set of acts $O_i(e_{i,t_i})$. In case the agent maximizes a rational preference relation, it is easy to notice $\mu^{-1}(C_{i,t_i'}(O_i(e_{i,t_i})))$ coincides with the set B_{i,t_i} considered in the main body of the paper. We moreover say σ is a solution to a mechanism γ whenever there exists $e \in E(\gamma)$ such that $\sigma \in R(e)$.

Let O_i^{f,t_i} denote the set of IAA acts that agent i can generate in the direct mechanism (f, T) by reporting $t_i' \in T_i$ when her opponents report truthfully. Incentive Compatibility

[22] This rules out, for example, preferences for truth-telling.

can then be generalized as in Saran (2011):

Definition 3 (Incentive Compatibility (IC))**.** *Let C_{i,t_i} be given. We say f satisfies IC for type $t_i \in T_i$ and $i \in I$ whenever $f(t_i, \cdot) \in C_{i,t_i}(O_i^{f,t_i})$. We say f is IC whenever it is IC for all $t_i \in T_i$ and $i \in I$.*

In other words, we require agents to choose the act associated with their type t_i when they expect their opponents to choose the acts associated to their types as well. In the case of BIC, this coincides with the set of acts maximizing expected utility in the choice set.

To derive our main result, it is useful to redefine WRC in terms of choice functions rather than utility maximization.

Definition 4 (Weak Choice Consistency (WCC))**.** *We say a solution concept $\mathcal{S}$ satisfies WCC for a class of mechanisms $\tilde{\Gamma} \subseteq \Gamma$ whenever for all $\gamma \in \tilde{\Gamma}$ $i \in I$, $t_i \in T_i$ there exists $\sigma \in \mathcal{S}(\gamma)$ such that $\mu(\sigma_i(t_i), \sigma_{-i}) \in C_{i,t_i}(X_i(\sigma_{-i}))$, where:*

$$X_i(\sigma_{-i}) = \{x_i \in \mathcal{A} : x_i = \mu(\sigma_i(t_i'), \sigma_{-i}), t_i' \in T_i\}$$

We say $\mathcal{S}$ satisfies WCC if it satisfies WCC for all $\gamma \in \Gamma$ such that $\mathcal{S}(\gamma) \neq \emptyset$.

It is immediate to extend Theorem 3 and, when F is a singleton, Theorem 1.

Theorem 7. *If F is implementable in WCC $\mathcal{S}$, then for all $i \in I$ and $t_i \in T_i$ there exists $f^{i,t_i} \in F$ that is IC for i and t_i. Conversely, if F is implementable and there exists $f^{i,t_i} \in F$ that is IC for i and t_i, then $\mathcal{S}$ is WCC for Γ^F.*

Notice both WRC and WCC implicitly assume a mild form of contraction consistency between choices in $O_i(\sigma_{-i})$ and in $X_i(\sigma_{-i})$. As a matter of fact, if $\sigma \in \mathcal{S}(\gamma)$ is such that $\mu(\sigma_i(t_i), \sigma_{-i}) \in C_{i,t_i}(O_i(\sigma_{-i}))$ and $X_i(\sigma_{-i}) \subseteq O_i(\sigma_{-i})$, WCC entails that $\mu(\sigma_i(t_i), \sigma_{-i}) \in C_{i,t_i}(X_i(\sigma_{-i}))$.

This implicit assumption means it is not as easy to provide a sufficient condition for WCC as it was for WRC. As a matter of fact, let us say $\mathcal{S}$ is Choice Consistent (CC) for mechanism γ whenever there exist $e, e' \in E(\gamma)$ and $\sigma \in R(e)$ such that $(\sigma_i, e_{i,t_i}) \in R(e')$. While this entails $\mu(\sigma_i(t_i), e_{i,t_i}) \in C_{i,t_i}(O_i(e_{i,t_i}))$, this is not enough to establish WCC unless we exclude the possibility that $\mu(\sigma_i(t_i), e_{i,t_i}) \notin C_{i,t_i}(X_i(e_{i,t_i}))$.

Sen (1971) provides an example of a class of choice functions ruling out such a possibility. We say a choice function C_{i,t_i} satisfies Independence of Irrelevant Alternatives (IIA) whenever for all $X, Y \subseteq \mathcal{A}$:

$$C_{i,t_i}(X) \subseteq Y \subseteq X \implies C_{i,t_i}(X) \subseteq C_{i,t_i}(Y)$$

It is then easy to see that IIA entails that if $\mathcal{S}(\gamma)$ is CC, it will be WCC as $\mu(\sigma_i(t_i), e_{i,t_i}) \in C_{i,t_i}(O_i(e_{i,t_i}))$ and:

$$C_{i,t_i}(O_i(e_{i,t_i})) \subseteq X_i(e_{i,t_i}) \subseteq O_i(e_{i,t_i}) \implies C_{i,t_i}(O_i(e_{i,t_i})) \subseteq C_{i,t_i}(O_i^{f,t_i})$$

We can then provide the following corollary to Theorem 7.

Corollary 4. *If F is implementable in CC $\mathcal{S}$ and C_{i,t_i} satisfies IIA, then there exists $f \in F$ that is IC for type t_i and agent i.*

B.1.1 Examples

It is clear any choice function that mandates agents to pick the best (or worst) alternative in their choice set according to total order $\succeq_{i,t_i}$ satisfies IIA. For example, if:

$$C_{i,t_i}(X) = \arg\max_{\succeq_{i,t_i}} X$$

Then $C_{i,t_i}(X) \subseteq Y$ directly entails:

$$\arg\max_{\succeq_{i,t_i}} X = \arg\max_{\succeq_{i,t_i}} Y$$

It is not necessary $\succeq_{i,t_i}$ to be generated by vNM preferences over lotteries: the model above can easily accommodate non-expected utility models as Prospect Theory (Kahneman and Tversky, 1979) as long as the ordering over lotteries does not depend on the choice set they are chosen from.

Models of satisficing (Simon, 1955) satisfy IIA as well. For example, suppose we require the alternatives chosen to be at least as good as a type- and agent- dependent benchmark alternative k_{i,t_i}:

$$C_{i,t_i}(X) = \{a \in X : a \succeq_{i,t_i} k_{i,t_i} \text{ where } k_{i,t_i} \in A\}$$

It is clear this model satisfies IIA: as $C_{i,t_i}(Y)$ consists of all the alternatives in Y that are weakly preferred to the benchmark k_{i,t_i}, $C_{i,t_i}(X) \subseteq Y$ implies $C_{i,t_i}(X) = C_{i,t_i}(Y)$. A similar argument applies for ϵ-BNE in the spirit of Radner (1980), which relaxes the definition of Cournot-Nash Equilibrium to allow agents to choose any action that grants them a payoff within ϵ of the maximum one.

As a counterexample, choice functions displaying attraction or compromise effects (e.g., the hiring model in de Clippel (2014)) do not fall in general into this category, as modifying the choice set can potentially affect agents' choices. Similarly, stochastic choice models *à la* (Luce, 1959) do not always satisfy this condition. This is due mainly to the fact the

support of the lotteries prescribed by the associated choice functions tends to extend beyond O_i^{f,t_i}. For example, in Quantal Response Equilibrium (QRE) (McKelvey and Palfrey, 1995) the probability each agent chooses action s_i depends negatively on the expected utility associated to each alternative action $s_i' \neq s_i$.

We can also consider setups that do not rely on the assumption utility is independent of i's expectations or the expectations of her opponents. For example, consider the following adaptation of the model of fairness equilibrium from Rabin (1993) to games of incomplete information. We will denote as $\pi_i(\mu(\sigma), t)$ the "material payoff" i derives from the outcome associated to profile σ, i's beliefs on how kind $j \neq i$ is being to her as $\tilde{\kappa}_j(e_i, e_j, t)$ and player i's kindness towards her opponents as $\kappa_i(\sigma_i, e_i, t)$. We can then write i's expected utility function as:

$$\bar{u}_i(s_i, t_i, e) = \int_{T_{-i}} \pi(\mu(s_i, \sigma_{-i}(t_{-i})), t) p_i(t_{-i}|t_i) + \sum_{j \neq i} \int_{T_{-i}} \tilde{\kappa}_j(e_i, e_j, t) \left[1 + \kappa_i(s_i, e_j, t)\right] \, dp_i(t_{-i}|t_i)$$

The following theory of behavior can then capture the set of equilibria in the model:

$$E^{ED}(\gamma) = \left\{ e \in \mathcal{E}(\gamma) : \exists \sigma \in \Sigma \text{ s.t. } e_i = \sigma_{-i}, \sigma_i(t_i) \in \arg \max_{s_i' \in \Delta(S_i)} \bar{u}_i(s_i, t_i, e) \text{ for all } i \in I \right\}$$

$$R^{ED}(e) = \{\sigma \in \Sigma : \sigma_{-i} = e_i\}$$

This definition implies expectations are rational, and thus that $\mathcal{S}^{ED}$ is CC. However, Incentive Compatibility does not necessarily hold: as agents' preferences may depend on the menu of acts that are available to them and their opponents, more information than the one conveyed by the SCF is needed to establish whether they would like to mimic a different type or not.

B.2 Epistemic Foundations

To better appreciate how restrictive the conditions for necessity of BIC are, this section provides an epistemic justification for (Weak) Response Consistency. The bulk of the argument is based on the one Dekel et al. (2007) propose for the characterization of interim correlated rationalizability. RC obtains whenever each type of each agent is rational, knows the type space and that she could successfully mimic other types. This makes overall requirements for RC rather weak, and the related class of solution concepts rather large.

Let γ be fixed, T be finite, and let $P_i(t_i') = \{t \in T : p_i(t_{-i}|t_i') > 0\}$. We moreover denote as $\mathcal{S}(\gamma, t)$ the set of profiles s such that there exists $\sigma \in \mathcal{S}(\gamma)$ such that $\sigma(t) = s$.

Let H_i be a finite set of epistemic types for player i, and let $H = \times_{i \in I} H_i$[23]. An epistemic model specifies:

- functions $\phi_i : H_i \to \Delta(H_{-i})$, mapping i's epistemic type into a belief over others' epistemic types
- i's epistemic strategy $\eta_i : H_i \to \Delta(S_i)$, assigning a distribution over actions to each epistemic type
- i's standard type $\tau_i : H_i \to T_i$

An epistemic model is then a tuple $(H_i, \phi_i, \eta_i, \tau_i)_{i \in I}$, with state space H. For given ϕ_i, denote the joint distribution over opponents' actions and standard types:

$$\lambda_i(h_i)[s_{-i}, t_{-i}] = \int_{H_{-i}} \eta_{-i}(h'_{-i})[s_{-i}] \mathbf{1}_{\{\tau_{-i}(h'_{-i}) = t_{-i})\}}(h'_{-i})\, d\phi_i(h_i)[h'_{-i}]$$

For all $s_{-i} \in \Delta(S_{-i})$, let $\lambda_i(h_i, t_{-i})[s_{-i}] = \lambda_i(h_i)[s_{-i}, t_{-i}]$ denote the probability profile s_{-i} is played conditional on type profile t_{-i}. We then define the event in which i's beliefs over her opponents' standard types are consistent with ϕ_i as:

$$W_i = \{h \in H : \lambda_i(h_i)[S_{-i}, \tau_{-i}(h_{-i})] = p_i(\tau_{-i}(h_{-i}) | \tau_i(h_i))\}$$

Notice $h \in W_i$ implies $\lambda_i(h_i, \cdot) : T_{-i} \to \Delta_{-i}(S_{-i})$, so that $\lambda_i(h_i, \cdot) \in \Sigma_{-i}$ is a properly defined expectation. Let RAT_i be the set of states such that, conditional on beliefs $\phi(h_i)$, playing the mixed action associated to one's own epistemic type h_i yields higher expected utility than the one associated with any other epistemic type h'_i:

$$RAT_i = \left\{ h \in H : \eta_i(h_i) \in \arg\max_{h'_i \in H_i} \int_{H_{-i}} u_i(\mu(\eta(h'_i, h_{-i})))\, d\phi_i(h_i)[h_{-i}] \right\}$$

We then define two more events. The first event can be interpreted as the set of epistemic states such that i's playing the action associated with her epistemic type will be a solution to the mechanism if her opponents do the same:

$$TT_i = \{h \in H : (\eta_i(h_i), \lambda_i(h_i, \tau_{-i}(h_{-i}))) \cap \mathcal{S}(\gamma, \tau(h)) \neq \emptyset\}$$

The second event can be interpreted as the set of epistemic states h such that i playing the action associated with an epistemic type different from her own will be a solution to the mechanism when her opponents play the action associated with their epistemic type:

$$PM_i = \{h \in H : (\eta_i(\tau_i^{-1}(t'_i)), \lambda_i(h_i, \tau_{-i}(h_{-i}))) \cap \mathcal{S}(\gamma, (t'_i, \tau_{-i}(h_{-i})) \neq \emptyset, t'_i \neq \tau_i(h_i)\}$$

[23]Finiteness of S, T and H is not necessary for the argument, but it simplifies the exposition by avoiding the use of measure-theoretic notation.

This drives necessity of (partial) BIC for full implementation: if type t_i can get away with mimicking any other type t_i', it will become necessary to provide her incentives not to do so by choosing a BIC SCF f. Formally, for any $t_i' \in T_i$, $\eta_i(\tau_i^{-1}(t_i'))$ represents the set of all actions played by epistemic types with standard type t_i'. So $(\eta_i(\tau_i^{-1}(t_i')), \lambda_i(h_i, \tau_{-i}(h_{-i}))$ corresponds to all distributions over actions i can induce by mimicking some type $t_i' \neq \tau_i(h_i)$, conditional on her expectations about the actions of her opponents. The requirement that at least one profile in $(\eta_i(\tau_i^{-1}(t_i')), \lambda_i(h_i, \tau_{-i}(h_{-i}))$ belongs to $\mathcal{S}(t_i', \tau_{-i}(h_{-i})$ can instead be interpreted by considering that, in order to successfully mimic a different type, t_i must also induce the planner into thinking that the profile of actions she observes comes from type profile $(t_i', \tau_{-i}(h_{-i}))$ rather than $(t_i, \tau_{-i}(h_{-i}))$.

Finally, we denote the intersection of the last two events as:

$$SOL_i = TT_i \cap PM_i$$

This leads to a second possible interpretation of the two events above: as in Zambrano (2008), we can take them as meaning the action η_i prescribes for i, together with her expectations, will form a solution to the game. Zambrano (2008) uses mutual knowledge of this event to characterize the set of correlated rationalizable action profiles. It is important to notice the characterization in Zambrano (2008), differently from others, relies only on *mutual* rather than *common* knowledge.

We can then prove that WRC is almost characterized by the existence of an epistemic type for each type $t_i \in T_i$ that is rational and knows the standard type space and that mimicking is possible[24].

Theorem 8. *The two following statements are equivalent:*

1. *there exists an epistemic model such that for all $i \in I$ and $t_i \in T_i$ there exists $h^* \in RAT_i \cap K_i(W_i \cap SOL_i)$ with $\tau_i(h_i^*) = t_i$*

2. *for all $i \in I$ and $t_i \in T_i$ there exists $\sigma^{i,t_i} \in \Sigma$ such that $\sigma^{i,t_i}(\tilde{t}) \in \mathcal{S}(\gamma, t)$ for all $\tilde{t} \in P_i(t_i)$ and for all $t_i' \in T_i$:*

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

[24]The property discussed in this section is weaker than WRC, as the fact that for all $t \in P_i(t_i)$ there exists $\tilde{\sigma}(t)$ such that $\tilde{\sigma}(t) = \sigma(t)$ does not generally imply $\sigma^{i,t_i} \in \mathcal{S}(\gamma)$. Such a gap is immaterial for the results of Section 4, but it would require the use of heavier notation.

We can apply a similar argument to RC in order to appreciate how different it is from WRC. The main difference will be due to the fact utility maximization does not characterize the set of responses anymore, and that such a set may depend on the full profile of expectations e. For any profile $\tilde{\sigma}_{-i} \in \Sigma_{-i}$, let $R_{i,t_i}(\tilde{\sigma}_{-i})$ denote the set of profiles σ that are a response to an expectation profile e consistent with type t_i of agent i expecting her opponents to play $\tilde{\sigma}_{-i}$. Formally:

$$R_{i,t_i}(\tilde{\sigma}_{-i}) = \{s_i \in \Delta(S_i) : s_i = \sigma(t) \text{ for } \sigma \in R(e) \text{ s.t. } e_{i,t_i} = \tilde{\sigma}_{-i}\}$$

We can then define:

$$RAT_i^* = \{h \in H : \eta_i(h_i) \in R_{i,\tau_i(h_i)}(\lambda_i(h_i, \cdot))\}$$

We can then characterize RC as follows.

Theorem 9. *The two following statements are equivalent:*

1. *there exists an epistemic model such that for all $i \in I$ and $t_i \in T_i$ there exists $h^* \in RAT_i^* \cap K_i(W_i \cap SOL_i)$ with $\tau_i(h_i^*) = t_i$*
2. *for all $i \in I$ and $t_i \in T_i$ there exists $e \in E(\gamma)$ and $\sigma \in \Sigma$ such that $\sigma_i(t_i) \in R_{i,t_i}(e_{t_i})$ and $(\sigma_i, e_{i,t_i})(\tilde{t}) \in \mathcal{S}(\gamma, \tilde{t})$ for all $\tilde{t} \in P_i(t_i)$*

Therefore, the main difference between the two conditions does not lie in what agents know but rather in the assumptions made on the way they respond to their expectations. In particular, it highlights the central role played by knowledge of SOL_i in determining whether a solution concept is WRC or RC. Cursed equilibrium in Section 5 is a case in point. In that example, even if in all solutions each type of each agent plays a pure strategy, $\bar{\sigma}_{-i}$ assigns equal probability to both actions. Thus $(\sigma_i(t_i), \bar{\sigma}_{-i}(t_{-i}))$ fails to be a solution of the mechanism for any type profile t.

B.3 Proofs for Appendix B

Proof of Theorem 7. Suppose F is implementable in WCC $\mathcal{S}$ via mechanism $\gamma = (\mu, S)$ with $\mathcal{S}(\gamma) \neq \emptyset$. Then for each $i \in I$ and $t_i \in T_i$ there exists $\sigma \in \mathcal{S}(\gamma)$ such that $\mu(\sigma_i(t_i), \sigma_{-i}) \in C_{i,t_i}(X_i)$, where:

$$X_i = \{x_i \in \mathcal{A} : x_i = \mu(\sigma_i(t_i'), \sigma_{-i}), t_i' \in T_i\}$$

As $\sigma \in \mathcal{S}(\gamma)$, implementability of F yields $\mu(\sigma) = f$ for some $f \in F$. Then:

$$X_i = \{x_i \in \mathcal{A} : x_i = \mu(\sigma_i(t_i'), \sigma_{-i}) = f(t_i', \cdot), t_i' \in T_i\} = O_i^{f,t_i}$$

Then $f(t_i, \cdot) = \mu(\sigma_i(t_i), \sigma_{-i}) \in C_{i,t_i}(X_i) = C_{i,t_i}(O_i^{f,t_i})$, which is enough to prove $f \in F$ is BIC for agent i and type t_i.

As for the converse, suppose F is implementable in $\mathcal{S}$ via mechanism γ and that for all $i \in I$ and $t_i \in T_i$ there exists $f \in F$ that is IC for agent i and type t_i. As f is IC for i and t_i, $f(t_i, \cdot) \in O_i^{f,t_i}$. Moreover, as F is implementable, there exists a solution $\sigma \in \mathcal{S}(\gamma)$ such that $f = \mu(\sigma)$. Then:

$$O_i^{f,t_i} = \{x_i \in \mathcal{A} : x_i = f(t_i', \cdot) = \mu(\sigma_i(t_i), \sigma_{-i}), t_i' \in T_i\} = X_i$$

As our initial choice of i and t_i was arbitrary, this concludes the proof. □

Proof of Theorem 8. $1 \implies 2$: Let $h^* \in RAT_i \cap K_i(W_i \cap SOL_i)$. We invoke the following Lemma, which is proved separately.

Lemma 1. *If $h^* \in K_i(W_i \cap SOL_i)$, there exists $z_i : T \to H_i$ such that $\tau_i(z_i(t)) = t_i$ for all $t_i \in T_i$ and $(\eta_i(z_i(t)), \lambda_i(h_i^*, t_{-i})) \in \mathcal{S}(\gamma, t)$ for all $t \in P_i(\tau_i(h_i^*))$.*

Set then $\sigma(t) = (\eta_i(z_i(t)), \lambda_i(h_i^*, t_{-i}))$ for all $t \in T$. Then $h^* \in RAT_i$ yields that for all $h_i' \in H_i$:

$$\int_{H_{-i}} u_i(\mu(\eta(h_i^*, h_{-i}))\, d\phi_i(h_i^*)[h_{-i}] \geq \int_{H_{-i}} u_i(\mu(\eta(h_i', h_{-i})))\, d\phi_i(h_i^*)[h_{-i}]$$

This entails that for all $t_i' \in T_i$:

$$\int_{H_{-i}} u_i(\mu(\eta(h_i^*, h_{-i}))\, d\phi_i(h_i^*)[h_{-i}] \geq \int_{H_{-i}} u_i(\mu(\eta(z_i(t_i', t_{-i}), h_{-i})))\, d\phi_i(h_i^*)[h_{-i}]$$

By Fubini-Tonelli and our choice of σ we can rewrite the inequality above as:

$$\int_{T_{-i}} u_i(\mu(\sigma(\tau_i(h_i^*), t_{-i})), t)\, dp_i(t_{-i}|t_i) \geq \int_{T_{-i}} u_i(\mu(\sigma(t_i', t_{-i})), t)\, dp_i(t_{-i}|t_i)$$

This concludes the proof.

$2 \implies 1$: Let H_j consist of all pairs (s_j, t_j) such that s_j is equal to $\sigma_j^{j,t_j}(t)$ for some $t_j \in T_j$[25]. Let $\eta_j(h_j) = \eta_j(s_j, t_j) = s_j$ and $\tau_j = \tau_j(s_j, t_j) = t_j$. For all $j \in I$ and $h_j \in H_j$, let:

$$\phi_j(h_j)[h_{-j}] = \phi_j(s_j, t_j)[(s_{-j}, t_{-j})] = \sigma_{-j}^{j,t_j}(t_{-j})[s_{-j}]p_j(t_{-j}|t_j)$$

We now show that this epistemic models is such that for all $i \in I$ and $t_i \in T_i$ there exists $h^* \in RAT_i \cap K_i(W_i \cap SOL_i)$ with $\tau_i(h_i^*) = t_i$.

[25] As T is finite, this entails H is finite as well.

Fix now any $i \in I$ and $t_i \in T_i$, and consider any h^* such that $h_i^* = (\sigma_i^{i,t_i}(t_i), t_i)$. By our choice of τ, it is immediate to notice $\tau_i(h_i^*) = t_i$.

We first show $h^* \in K_i(W_i)$. As $\sigma_{-i}^{i,t_i}(t_{-i})[S_{-i}] = 1$ by definition of Σ_{-i}, for all $t_{-i} \in T_{-i}$:

$$\int_{H_{-i}} \mathbf{1}_{\{\tau_{-i}(h_{-i})=t_{-i}\}}(h_{-i})\, d\phi_i(h_i)[h_{-i}] = \phi_i(s_i, t_i)[(S_{-i}, t_{-i})] = p_i(t_{-i}|t_i)$$

Moreover, $h^* \in RAT_i$. Notice first that for all $h_i' \in H_i$, there exists $t_i' \in T_i$ such that $\eta(h_i') = \sigma_i^{i,t_i}(t_i')$ by our definition of H_i. As by assumption we have:

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

We can use Fubini-Tonelli to rewrite the inequality above as:

$$\int_{H_{-i}} u_i(\mu(\eta(h_i^*, h_{-i}))\, d\phi_i(h_i^*)[h_{-i}] \geq \int_{H_{-i}} u_i(\mu(\eta(h_i', h_{-i}))\, d\phi_i(h_i^*)[h_{-i}]$$

Which is enough to prove $h^* \in RAT_i$.

Suppose now for the sake of contradiction that $\phi_i(h_i^*)[SOL_i] < 1$, so that $h^* \notin K_i(SOL_i)$. Then there exists an epistemic state h_{-i} with $\phi_i(h_i^*)[h_{-i}] > 0$ such that either:

$$(\eta_i(h_i^*), \lambda_i(h_i^*, \tau_{-i}(h_{-i})) \notin \mathcal{S}(\gamma, \tau(h_i^*, h_{-i}))$$

Or for some standard type t_i' we have that for all h_i with $\tau_i(h_i)$:

$$(\eta_i(h_i), \lambda_i(h_i^*, \tau_{-i}(h_{-i})) \notin \mathcal{S}(\gamma, (t_i', \tau_{-i}(h_{-i}))$$

In either case, $\phi_i(h_i^*)[h_{-i}] > 0$ implies $p_i(\tau_{-i}(h_{-i})|t_i) > 0$ and thus either:

$$(\sigma_i^{i,t_i}(t_i), \lambda_i(h_i^*, \tau_{-i}(h_{-i})) = \sigma^{i,t_i}(t_i, \tau_{-i}(h_{-i})) \notin \mathcal{S}(\gamma, (t_i, \tau_{-i}(h_{-i}))$$

Or:

$$(\sigma_i^{i,t_i}(t_i'), \lambda_i(h_i^*, \tau_{-i}(h_{-i})) = \sigma^{i,t_i}(t_i', \tau_{-i}(h_{-i})) \notin \mathcal{S}(\gamma, (t_i', \tau_{-i}(h_{-i}))$$

This contradicts our premises, concluding the proof. □

Proof of Theorem 9. $1 \implies 2$: Let $h^* \in RAT_i^* \cap K_i(W_i \cap SOL_i)$. We invoke again Lemma 1 to argue there exists $z_i : T \to H_i$ such that $\tau_i(z_i(t)) = t_i$ for all $t_i \in T_i$ and $(\eta_i(z_i(t)), \lambda_i(h_i^*, t_{-i})) \in \mathcal{S}(\gamma, t)$ for all $t \in P_i(\tau_i(h_i^*))$.

Consider now profile $(\eta_i(z_i(t)), \lambda_i(h_i^*, t_{-i}))$ for all $t_i \in T_i$. From $h^* \in RAT_i^*$ there exists $e \in E(\gamma)$ and $\sigma \in R(e)$ such that $e_{i,t_i} = \lambda_i(h_i, \cdot)$ and $\eta_i(h_i^*) = \sigma_i(\tau_i(h_i^*))$. We can then rewrite:

$$(\sigma_i, e_{i,t_i})(t) = (\eta_i(z_i(t)), \lambda_i(h_i^*, t_{-i}))$$

By construction of z_i, for all $t \in P_i(\tau_i(h_i^*))$:

$$(\sigma_i, e_{i,t_i})(t) = (\eta_i(z_i(t)), \lambda_i(h_i^*, t_{-i})) \in \mathcal{S}(\gamma, t)$$

This concludes the proof.

$2 \implies 1$: The proof is analogous to the one of Theorem 8, except for the fact we now have to show $h^* \in RAT_i^*$. This follows easily as $\sigma \in R(e)$, $e_{i,t_i} = \lambda_i(h_i^*, \cdot)$ and $\eta_i(h_i^*) = \sigma_i(t_i)$. □

Proof of Lemma 1. We invoke the following Lemma, proved below:

Lemma 2. *If $p_i(t_{-i}|\tau_i(h_i^*)) > 0$ for $t_{-i} \in T_{-i}$, there exists $h_{-i} \in H_{-i}$ such that $\tau_{-i}(h_{-i}) = t_{-i}$ and $(h_i^*, h_{-i}) \in SOL_i$.*

Let $t \in T$ be such that $p_i(t_{-i}|\tau_i(h_i^*)) > 0$, and let $(h_i^*, h_{-i}) \in SOL_i$ be such that $\tau_{-i}(h_{-i}) = t_{-i}$ (we know such h_{-i} exists by Lemma 2).

As $(h_i^*, h_{-i}) \in SOL_i$, for all $t_i \in T_i$ there exists $h_i \in \tau_i^{-1}(t_i)$ such that:

$$(\eta_i(h_i), \lambda_i(h_i^*, t_{-i})) \cap \mathcal{S}(\gamma, t) \neq \emptyset$$

For all $t \in P_i(\tau_i(h_i^*))$ with $t_i \neq \tau_i(h_i^*)$, set then $z_i(t)$ equal to such h_i. By an analogous argument, for all $t \in P_i(\tau_i(h_i^*))$ with $t_i = \tau_i(h_i^*)$, set $z(t) = h_i^*$, and let $z(t)$ be any $h_i \in \tau_i^{-1}$ for $t \notin P_i(\tau_i(h_i^*))$.

It is clear $\tau_i(z_i(t)) = t_i$, as $z_i(\tau_i(h_i^*), t_{-i}) = h_i^*$ and $z_i(t) \in \tau_i^{-1}(t_i)$ for $t_i \neq \tau_i(h_i^*)$. Moreover, the argument above implies that for all $t \in T$:

$$(\eta_i(z(t)), \lambda_i(h_i^*, t_{-i})) \cap \mathcal{S}(\gamma, t) \neq \emptyset$$

Concluding the proof. □

Proof of Lemma 2. Let $SOL_i(h_i^*) = \{h \in H : (h_i^*, h_{-i}) \in SOL_i\}$. By $K_i(SOL_i)$ we have:

$$\int_{SOL_i(h_i^*)} 1 \, d\phi_i(h_i^*)[h_{-i}] = 1$$

As T_{-i} is finite, we can break down the sum over all type profiles of i's opponents as:

$$\sum_{t_{-i} \in T_{-i}} \int_{SOL_i(h_i^*)} \mathbf{1}_{\{\tau_{-i}(h_{-i}) = t_{-i}\}}(h_{-i}) \, d\phi_i(h_i^*)[h_{-i}] = \int_{SOL_i(h_i^*)} 1 \, d\phi_i(h_i^*)[h_{-i}] = 1$$

By $K_i(W_i)$ it moreover holds:

$$p_i(t_{-i}|\tau_i(h_i^*)) = \int_H \mathbf{1}_{\{\tau_{-i}(h_{-i}) = t_{-i}\}}(h_{-i}) \, d\phi_i(h_i^*)[h_{-i}] \geq \int_{SOL_i(h_i^*)} \mathbf{1}_{\{\tau_{-i}(h_{-i}) = t_{-i}\}}(h_{-i}) \, d\phi_i(h_i^*)[h_{-i}]$$

Where the last inequality follows from $SOL_i(h_i^*) \subseteq H$. As the sum over types of both sides is one, from this inequality it follows that:

$$p_i(t_{-i}|\tau_i(h_i^*)) = \int_{SOL_i(h_i^*)} \mathbf{1}_{\{\tau_{-i}(h_{-i})=t_{-i}\}}(h_{-i})\, d\phi_i(h_i^*)[h_{-i}]$$

Then, if $p_i(t_{-i}|\tau_i(h_i^*)) > 0$, there exists at least one $h_{-i} \in H_{-i}$ such that $h_{-i} \in SOL_i(h_i^*) \subseteq SOL_i$ and $\tau_{-i}(h_{-i}) = t_{-i}$. □