The Revelation Principle without Rational Expectations

Giacomo Rubbini*

November 5, 2023

Most recent version

Abstract

The revelation principle states that it is without loss of generality to restrict attention to direct mechanisms and, consequently, that incentive compatibility is necessary for implementation in Bayesian Nash Equilibrium. This paper extends the discussion beyond Bayesian Nash Equilibrium by providing sufficient conditions on the solution concept that ensure any implementable social choice function can be implemented via a direct mechanism. These conditions do not generally imply incentive compatibility is necessary for implementation, as the class of solution concepts requiring incentive compatibility for implementation is characterized via a logically independent condition.

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 Pedro Dal Bó, Pietro Dall'Ara, Geoffroy De Clippel, Zeky Murra Anton, and Kareen Rozen for useful comments and suggestions. All errors are my own.

It might seem that problem of optimal auction design must be quite unmanageable, because there is no bound on the size or complexity of the strategy spaces which the seller may use in constructing the auction game. The basic insight which enables us to solve auction design problems is that there is really no loss of generality in considering only direct revelation mechanisms.

Roger Myerson (1981)

1 Introduction

The revelation principle is one of the most influential results in mechanism design, as it entails any social goal that is implementable in Bayesian Nash Equilibrium (BNE) can be implemented by simply asking agents to reveal their private information (i.e., their type). By relying only on these *direct revelation mechanisms*, the designer can avoid potentially complex and impractical indirect mechanisms without any loss of generality, significantly simplifying the analysis of many applied mechanism design problems.

For instance, in auction design problems (Myerson, 1981), the revelation principle entails we can restrict attention to mechanisms in which bidders submit their valuations to the seller, who then determines the winner of the object and each agent's payment as a function of the vector of bids. Likewise, in bilateral trading problems (for instance, Myerson and Satterthwaite 1983), we can focus on mechanisms in which the seller and the buyer submit their valuation for the object and probability of trade and transfers are determined according to the valuations reported.

As the mechanism designer can just ask agents to report their private information, it is crucial to incentivize them to do so truthfully. For BNE mechanism design, this means that truth-telling must be an equilibrium of the direct mechanism. This means Bayesian Incentive Compatibility is a necessary condition for implementation— that is, the social

goal must incentivize each agent to truthfully reveal her private information, as long as their opponents are telling the truth as well. It is still unclear whether the same holds true for solution concepts different from BNE.

This paper extends the revelation principle beyond BNE implementation, particularly to solution concepts that do not assume rational expectations. We identify a class of solution concepts for which the revelation principle holds via three conditions that are simple to interpret. Additionally, we characterize the class of solution concepts that require Bayesian Incentive Compatibility (BIC) for partial implementation. This second class is characterized by a property logically independent from the ones behind the revelation principle.

Theorem 1 in Section 3 provides three sufficient conditions on the solution concept that jointly imply it is without loss of generality to focus on direct mechanisms: Outcome Consequentialism (OC), Separation Invariance (SI), and Independence of Irrelevant Actions (IIA). Even if these conditions do not characterize the class of solution concepts of interest, the class of solution concepts satisfying all three provides considerable insight into the solution concept features that may be problematic when focusing on direct mechanisms only (see the discussion about Interim Correlated Rationalizability in Section 3.1.5 for an example).

OC is a mild condition, as it only requires the solution concept to be invariant to relabeling agents' actions—that is, the set of solutions depends on the consequences of agents' action profiles but not on the name of the actions themselves. SI requires instead that we can not knock off a solution to the mechanism by duplicating one or more of the actions an agent plays in that solution.[1] The third condition, IIA, requires that removing from a mechanism those (mixed) actions that are not played in a solution does not affect the existence of that solution.

IIA is the most restrictive condition out of the three as, if it does not hold, the planner

[1]It may happen that, in a solution, two types of an agent play the same (possibly mixed) action. Duplicating actions allows us to implement the same outcome while separating different types—that is, while ensuring different types play different actions.

may have to resort to mechanisms that are more complex than the direct mechanism. IIA ensures that the actions agents play in a solution contain all the information the planner needs to implement the social choice function of interest. For example, this is not the case for full implementation in BNE: actions that are not chosen in equilibrium provide additional information that the planner can use to knock off unwanted equilibria.

Corollary 1 moreover shows IIA is relatively more important than SI, as we can obtain a result similar to the revelation principle even without SI. This *quasi-revelation principle* entails the planner can ask agents just the information that is relevant to the goal the planner wants to implement. For example, consider two bidders who could value an object either \$0, \$50, or \$100. The revelation principle entails tells us the auctioneer can just ask each agent "How much is this object worth to you?". However, if the auctioneer wants to implement the same outcome regardless of whether one of the bidder's valuations is \$0 or \$50, some of this information is irrelevant. For instance, if the auctioneer does not want to sell the object below \$51, she does not need to discriminate agents with value \$0 from those with value \$50. The quasi-revelation principle then entails the planner can simply ask "Is the object more than 50 to you?" rather than "Is the object worth $0, 50$ or 100 to you?" as she would do with a direct mechanism.

Our results highlight that rational expectations are not necessary for the revelation principle to hold. In fact, other than Bayesian Nash Equilibrium and undominated Bayesian Nash Equilibrium, the class of solution concepts satisfying all three properties contains Cursed Equilibrium (Eyster and Rabin, 2005), and some variants of level-k reasoning models.

Theorem 1 shows that we can restrict attention to truth-telling in direct mechanisms, but it is silent on whether truth-telling will be an equilibrium of the mechanism—that is, on whether BIC is still necessary for implementation. Section 4 shows that BIC is necessary when all agents can infer the mechanism's solution the planner will select when implementing a social choice function—for instance, if we assume agents' expectations are consistent

with the planner maximizing known, strictly convex preferences.[2] The intuition behind the result follows Rubbini (2023), where it is established that BIC is necessary for full implementation of SCF if agents can correctly predict the mechanism's outcome. The rationale behind the result is, however, different: rather than following from the restrictiveness of full implementation, it follows here whenever agents can infer which solution the planner would like to implement.

Our result also suggests we pay particular caution in justifying the use of a partial implementation approach. A focus on partial implementation is justified by supposing, for example, that the planner may recommend agents to follow one particular solution of the game or that some solution (for example, truth-telling or risk-dominant behavior) is focal. This justification, however, may entail that agents can infer which solution of the mechanism will prevail and, subsequently, that BIC is necessary for partial implementation. de Clippel et al. (2019) highlight an issue of this kind in their Remark 2. In that case, to rationalize the strategies used to partially implement a non-BIC SCF, the planner needs to rely on agents' beliefs that are not consistent with the SCF the planner is implementing.

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

[2] For example, if the planner prefers an equal distribution of surplus generated from trade in the model of Myerson and Satterthwaite (1983).

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 $\sigma : \Omega \times T \to \Delta(S)$. We can interpret Ω as a set of mechanism-independent *strategic states* capturing those features of the strategic interaction that affect the mechanism's outcome but not the one the planner seeks to implement.

For instance, we can see the set of all possible level combinations in Kneeland (2022) as a set of strategic states: while the planner wants to condition the outcome produced by the mechanism on t only, actual play will depend on agents' level as well. Similarly, if we consider robust implementation (Bergemann and Morris, 2005), we can assume T is the set of agents' payoff types while Ω is the set of their belief types: while beliefs affect how agents play, the planner does not condition the outcome she wants to implement on agents' beliefs. To capture full implementation, we can also assume the set of strategic states to be the set of all equilibrium selection rules (Section 3.1.2).

We will denote the set of functions $\sigma : \Omega \times T \to \Delta(S)$ as Σ. For all $i \in I$, we will moreover let Σ_i (respectively, Σ_{-i}) denote the set of $\sigma_i : \Omega \times T_i \to \Delta(S_i)$ (respectively, $\sigma_{-i} : \Omega \times T_{-i} \to \Delta(S_{-i})$). For the rest of the paper, we will slightly abuse the notation above by considering $\mu(\sigma(\omega, t))$ to denote the lottery over A induced by $\sigma(\omega, t)$ under the outcome function μ.

[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.

Let us moreover assume A, T_i, and S_i are 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 are measurable functions.

3 Implementation via Direct Mechanisms

The revelation principle implies it is without loss of generality to restrict attention to direct mechanism when studying the implementation problem: for every social choice function, there exists then a solution of the direct mechanism that allows the planner to extract all the information she needs for implementation.

For partial implementation in Bayesian Nash Equilibrium (BNE), the revelation principle follows immediately from implementability, which implies truthful reporting is an equilibrium of the direct mechanism. It is unclear whether a similar result holds for different solution concepts and what properties the solution concept should possess for the revelation principle to hold.

We say solution concept $\mathcal{S}$ satisfies the *revelation principle* whenever all implementable social choice functions f are implementable via truthful reporting in the associated direct mechanism (f, T). We can capture this intuition with the following definition.

Definition 1 (Revelation Principle)**.** *Let $\tau : T \to T$ be the identity function. A solution concept $\mathcal{S}$ satisfies the revelation principle whenever any implementable SCF f is such that $\tau \in \mathcal{S}((f, T))$.*

This implies we can characterize the class of solution concepts such that the revelation principle holds as follows.

Remark 1. *$\mathcal{S}$ satisfies the revelation principle if and only if, for all γ and $\sigma \in \mathcal{S}(\gamma)$, $\tau \in \mathcal{S}((\mu \circ \sigma, T))$.*

In other words, the outcome associated with any solution σ of any mechanism γ can be replicated using a mechanism that uses as outcome function $\mu \circ \sigma$ and as action space T (a direct mechanism).

While the characterization of Remark 1 does not provide a lot of insight into the revelation principle, we can provide three more informative sufficient conditions. These conditions present a gap with the one that are necessary for the revelation principle to hold, but they capture three key insights about implementation in BNE via direct mechanisms:

- any information agents can convey through their actions can be equivalently conveyed by reporting their type (Outcome Consequentialism)
- the direct mechanism separates different types (Separation Invariance)
- each action is played by some type (Independence of Irrelevant Actions)

Even if these conditions do not characterize the class of solution concepts for which the revelation principle holds, checking whether a solution concept satisfies these properties can help in understanding why the revelation principle fails. See, for example, the discussion about Interim Correlated Rationalizability in Section 3.1.5: the fact IIA is violated is not enough to conclude the revelation principle does not hold, but it suggests actions that are not played in a solution could be an issue. This allows us to construct a counterexample that indeed shows that there exist SCFs that are implementable only with the help of indirect mechanisms.

We say a solution concept $\mathcal{S}$ is *outcome-consequential* (OC) if for all $\gamma, \tilde{\gamma}$ and bijective $\rho : S \to \tilde{S}$ such that $\tilde{\mu} \circ \rho = \mu$:

$$\sigma \in \mathcal{S}(\gamma) \implies \rho(\sigma) \in \mathcal{S}(\tilde{\gamma})$$

That is, if $\tilde{\gamma}$ is just the same as γ except for the fact all action profiles were relabeled with a different name, relabeling action of solution σ in $\mathcal{S}(\gamma)$ leads to a solution in $\mathcal{S}(\tilde{\gamma})$.

We conjecture this condition is satisfied by most solution concepts, as solutions typically depend only on the outcome agents can induce with their actions but not on the *name* of the actions themselves.[5]

For any two mechanisms γ and $\tilde{\gamma}$ such that $S \subseteq \tilde{S}$ and $\tilde{\mu}$ extends μ, let us say action $\tilde{s}_i \in \tilde{S}_i$ is a *duplicate* of mixed action $s_i \in \Delta(S_i)$ whenever $\tilde{\mu}(\tilde{s}) = \mu(s_i, \tilde{s}_{-i})$ for all $\tilde{s}_{-i} \in \tilde{S}_{-i}$. Let us denote as $D(s_i)$ the set of all actions that are duplicates of s_i. For any such γ and $\tilde{\gamma}$, let us moreover say a mechanism $\tilde{\gamma}$ is a *redundant extension* of γ if each $\tilde{s}_i \in \tilde{S}_i$ is a duplicate of a mixed action in $\Delta(S_i)$, for all $i \in I$.

We say a solution concept is *separation invariant* (SI) whenever duplicating an agent's action to separate different types does not knock off solutions to the mechanism. Formally, we require that for any γ, $\sigma \in \mathcal{S}(\gamma)$ and any redundant extension $\tilde{\gamma}$ of γ, there exists $\tilde{\sigma} \in \mathcal{S}(\tilde{\gamma})$ such that $\tilde{\sigma}_i(\omega, t) \neq \tilde{\sigma}_i(\omega, (t_i', t_{-i}))$ and $\tilde{\sigma}_i(\omega, t) \in D(\sigma_i(\omega, t))$ for all $i \in I$, $t_i, t_i' \in T_i$ and $t_{-i} \in T_{-i}$.

Finally, we say a solution concept $\mathcal{S}$ satisfies *independence of irrelevant actions* (IIA) whenever $\sigma \in \mathcal{S}(\gamma)$ implies there exists $\tilde{\sigma} \in \mathcal{S}(\tilde{\gamma})$, where $\tilde{S}_i = \sigma_i(\Omega, T)$ for all $i \in I$, $\tilde{\mu}$ is the restriction of μ to $\tilde{S}$, and $\tilde{\sigma}(\omega, t)[\sigma(\omega, t)] = 1$ for all $\omega \in \Omega$ and $t \in T$. That is, IIA is satisfied whenever the solution of a mechanism is not sustained from the presence of mixed action profiles that are never played in that solution.

IIA is the property with the tightest links with rational expectations. As a matter of fact, IIA implies the incentives a player has to report her own type (either in a direct mechanism or through her actions in an indirect one) do not depend on mixed action profiles that are played with zero probability: the information a player's strategy conveys to the planner does not depend on actions that are never played according to that strategy. This intuition is also similar to the one presented in Saran (2011), which shows the revelation principle holds only on the domain of menu independent preferences.

[5] A possible exception is those solution concepts in which we have a preference for truth-telling. However, such solution concepts are not usually well-defined for general mechanisms.

These three conditions are indeed sufficient to ensure the revelation principle holds.

Theorem 1. *If $\mathcal{S}$ satisfies OC, SI and IIA, then it satisfies the revelation principle.*

We relegate the proof of this result to the Appendix, and we give here only a sketch of the argument. Notice that as f is implementable, there exist a mechanism γ with a solution σ such that $\mu(\sigma) = f$. The first step of the proof involves removing from the action space of each player those actions that are never played in solution σ, generating a new (reduced) mechanism γ'. By IIA, the same distribution over actions generated by σ is a solution σ' of γ'. We then construct a second mechanism γ'' duplicating those actions that are played by more than one type to ensure each type plays a different action. SI then ensures there is a solution σ'' of this augmented mechanism that replicated the outcome of σ' (and thus of σ). Finally, we use OC to map each action profile in mechanism γ'' to the corresponding type profile in the direct mechanism.

Of the three conditions, OC and IIA are the most important to simplify a potentially complex indirect mechanism in a simpler direct one. Indeed, if SI does not hold, we can still construct something similar to a direct mechanism that does not separate different types, as pointed out in the following example.

Assume there are only two agents A and B, with three types for agent A (t_A, t'_A, t''_A) and one type for agent B (t_B, t'_B). Suppose it does not matter to the planner whether A is of type t_A or t'_A, as the SCF of interest prescribes the same outcome in either case: for instance, let $f(t_A, t_B) = f(t'_A, t_B) \neq f(t''_A, t_B)$ and $f(t_A, t'_B) = f(t'_A, t'_B) \neq f(t''_A, t'_B)$. A direct mechanism would ask each agent whether their type is t_A, t'_A or t''_A. However, some of the information extracted this way is irrelevant to the planner, as the distinction between type t_A and type t'_A is immaterial to implementation of f. The planner could instead ask to agent A a simpler question: "Is your type t''_A"? or, more explicitly, "Is your type in $\{t_A, t'_A\}$ or is it t''_A"?

We capture this intuition by defining the *quasi-direct mechanism* that extracts more

coarse information from the agent than a direct mechanism. Denote the quasi-direct mechanism associated to f as $(\hat{f}, \hat{T})$, where $\hat{T}_i$ is a partition of T_i such that for all $\hat{t}_i \in \hat{T}_i$:

$$t_i \in \hat{t}_i \text{ and } f(t'_i, t_{-i}) = f(t) \text{ for all } t_{-i} \in T_{-i} \implies t'_i \in \hat{t}_i$$

and $\hat{f}(\hat{t}) = f(t)$ for all $t \in \hat{T}$ and $\hat{t} \in \hat{T}$.

If SI does not hold, we can still construct a quasi-direct implementing mechanism to simplify the planner's task by first using IIA to reduce each agent's action space, and then using OC to map each of the resulting action profiles to the "coarse" type space $\hat{T}$.[6]

Corollary 1. *If $\mathcal{S}$ satisfies OC and IIA, any implementable SCF f is implementable via the associated quasi-direct mechanism $(\hat{f}, \hat{T})$.*

3.1 Examples

Of the three conditions presented above, IIA seems to be the most restrictive and the one closer to the assumption of equilibrium behavior. This intuition is confirmed by examining a few solution concept to check whether OC, SI and IIA hold.

A broad class of solution concepts satisfy all three: here we consider BNE, undominated BNE, cursed equilibrium, and level-k reasoning (when the planner is allowed to pick the anchor). IIA fails, instead, for Interim Correlated Rationalizability: even if an agent never plays a given action, this action may still rationalize some other agent's strategy. A similar intuition holds for level-k models in which the anchor is taken as exogenous. For example, de Clippel et al. (2019) assume the population of players does not feature any level-0: even if the actions prescribed by the anchor are never played by agents with a positive level, those actions may be the only ones rationalizing their play. Moreover, IIA is violated even if we consider full implementation in Bayesian Nash Equilibrium, as actions that are never

[6] We omit the full proof for this result, as it follows almost immediately from the proof of Theorem 1.

played may be necessary to knock off unwanted equilibria.

Unless otherwise specified, we assume Ω is a singleton for all the examples in this section. This means there are no features of the strategic interaction that affect play in the mechanism but do not outcome the planner seeks to implement in each state $t \in T$.

3.1.1 BNE

Let us say profile σ is a BNE of γ ($\sigma \mathcal{S}^{BNE}(\gamma)$) whenever there exists $\omega \in \Omega$ such that for all $i \in I$, $t_i \in T_i$ and $s_i \in S_i$:

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

Let $\tilde{\gamma} = (\tilde{\mu}, \tilde{S})$ be such that there exists a bijection $\rho : S \to \tilde{S}$ with $\mu = \tilde{\mu} \circ \rho$. Then, the set of inequalities above implies $\rho(\sigma)$ is an equilibrium of $\mathcal{S}^{BNE}(\tilde{\gamma})$ and OC holds.

To see BNE satisfies SI, notice by the construction of mechanism $\tilde{\gamma}$ we have $\mu(s_j, \sigma_{-j}) = \mu(s_j, \tilde{\sigma}_{-j})$ for all $j \in I$ and $s_j \in S_j$ and $\mu(\tilde{s}_i, \tilde{\sigma}_{-i}) = \mu(\sigma_i(\omega, t_i), \tilde{\sigma}_{-i}))$. It therefore follows from the definition of BNE that for all $i \in I$ and $\tilde{s}_i \in \tilde{S}_i$:

$$\int_{T_{-i}} u_i(\mu(\tilde{\sigma}(t)), t)\, dp_i(t_{-i}|t_i) \geq \int_{T_{-i}} u_i(\mu(\tilde{s}_i, \tilde{\sigma}_{-i}(\omega, t_{-i})), t)\, dp_i(t_{-i}|t_i)$$

Thus completing the proof.

BNE satisfies IIA as well by a similar argument. Let $\tilde{S}_i = \sigma_i(\Omega, T)$ and $\tilde{\sigma} : \Omega \times T \to \tilde{S}$ be such that $\tilde{\sigma}_i(t_i)[\sigma_i(\omega, t_i)] = 1$ for all $i \in I$, $t_i \in T_i$ and $s_i \in S_i$. Then by the definition of a BNE it follows again that for all $i \in I$, $t_i \in T_i$ and $s_i' \in \tilde{S}_i \subseteq S_i$:

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

Concluding the proof.

3.1.2 Full Implementation in BNE

While not the focus of the current paper, we can interpret Ω is a sense broad enough to capture full implementation in pure BNE within the current framework.[7]

Let Ω be the set of all functions ξ mapping each mechanism γ into one of its BNEs and let $\sigma \in \mathcal{S}^{FI}(\gamma)$ whenever $\sigma(\xi, \cdot) = \xi(\gamma)$ for all possible ξ. In this sense, we can interpret ξ as an *equilibrium selection rule* telling us which of the possible equilibria of the strategic interaction is actually realized. This captures the basic intuition of full implementation that the SCF should be implemented by the mechanism regardless of what equilibrium of the mechanism we consider. Therefore, the class of SCFs implementable in $\mathcal{S}$ coincides with the class of SCFs fully implementable in BNE.

$\mathcal{S}$ does not generally satisfy IIA: hence, the revelation principle does not hold for full BNE implementation.[8] This is due to the fact that, with respect to partial implementation, the planner needs more information to rule out the possibility that they are all pretending to be of a different type. As a matter of fact, IIA is akin to the conditions it is sufficient to impose on the equilibrium strategies of an implementing mechanism to make sure that the revelation principle holds (Postlewaite and Schmeidler (1986), Propositions 5-6).

To keep the discussion simple, let me focus on pure BNE implementation for the rest of this section.[9] Jackson (1991) proves full implementation requires f to satisfy Bayesian Monotonicity — that is, it requires there exists a *reward function* the planner can use to reward any agent informing her that the agents are mimicking a different type profile. The implementing mechanism Jackson (1991) construct, however, is an indirect one.

It is possible to characterize the class of SCFs implementable via direct mechanisms by

[7] We consider only pure BNE for ease of exposition, similar arguments carry through to the more general setup in Kunimoto (2019).

[8] Example 1 of Postlewaite and Schmeidler (1986) is a case point: to knock off unwanted (and possibly salient) equilibria, it is necessary to resort to indirect mechanisms.

[9] The argument can easily be extended via the results from Serrano and Vohra (2010) and Kunimoto (2019).

introducing a new condition. Let us say a SCF f satisfies *direct monotonicity* whenever for all deceptions $\alpha : T \to T$ such that $f \neq f \circ \alpha$ then there exists $i \in I$ and $t_i, t_i' \in T_i$ such that:

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

And for all $i \in I$ and $t_i \in T_i$:

$$\int_{T_{-i}} u_i(f(t), t)\, dp_i(t_{-i}|t_i) \geq \int_{T_{-i}} u_i(f(\alpha_i(t_i), t_{-i}), t)\, dp_i(t_{-i}|t_i)$$

The class of SCFs satisfying direct monotonicity is a subset of those satisfying Bayesian Monotonicity, as it is tantamount to imposing additional requirements on the "reward function".[10]

Theorem 2. *A SCF f is fully implementable in pure BNE via a direct mechanism if and only if it is BIC and directly monotonic.*

Direct monotonicity entails each equilibrium of the implementing mechanism contains all the information the planner needs to implement the SCF f of interest. For any state t state such that $f(t) \neq f \circ \alpha(t)$, the planner can infer the state is $\alpha(t)$ rather than t from observing $\sigma(\alpha(t))$ and not observing $\sigma_i(t_i')$. If the planner needed to not to observe some action $s_i' \notin \sigma_i(T_i)$ to reach the same conclusion, it would instead mean σ does not contain enough information for the planner to implement f, leading to a violation of IIA.

3.1.3 Undominated BNE

We say $\sigma \in \mathcal{S}^{UBE}(\gamma)$ is an undominated BNE of γ whenever it is a BNE and σ is not weakly dominated for any $i \in I$ and $t_i \in T_i$ (Palfrey and Srivastava, 1989). That is, σ is such that

[10] Example 1 of Postlewaite and Schmeidler (1986) indeed establishes inclusion is strict, as there exist SCFs that are fully implementable in pure BNE only through indirect mechanisms.

there is no $i \in I$, $t_i \in T_i$ and $\sigma_i' \in \Sigma_i$ such that:

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

To prove undominated BNE satisfies OC, we just follow the same steps as for BNE. Undominated BNE satisfies SI as well. Construct mechanism $\tilde{\gamma}$ and solution $\tilde{\sigma}$ as per the definition of SI. By the argument in the previous sections, $\tilde{\sigma}$ is a BNE. Moreover, as the set of payoffs each type of each agent can achieve is exactly the same in both γ and $\tilde{\gamma}$, the fact σ is undominated implies $\tilde{\sigma}$ is undominated.

Lastly, we prove undominated BNE satisfies IIA as well. Construct mechanism $\tilde{\gamma}$ and solution $\tilde{\sigma}$ as per the definition of IIA. Again, $\tilde{\sigma}$ will be a BNE of $\tilde{\gamma}$ by the argument in the previous sections. Suppose now, for the sake of contradiction, $\tilde{\gamma}$ is weakly dominated. As $\tilde{\sigma}$ yields exactly the same payoff as σ for all $i \in I$ and $t_i \in T_i$ and $\tilde{S} \subseteq S$, the fact σ is undominated implies $\tilde{\sigma}$ is undominated as well.

3.1.4 Cursed Equilibrium

We say a profile σ is a cursed equilibrium of mechanism γ whenever for all $i \in I$, $t_i \in T_i$ and $s_i' \in S_i$:

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

Where for $\chi \in [0, 1]$:

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

We can first notice CE satisfies SI. Suppose σ is a cursed equilibrium of γ, i.e. that it satisfies the inequality above for all $i \in I$, $t_i \in T_i$ and $s_i' \in S_i$. As above, by construction of mechanism $\tilde{\gamma}$ we have $\mu(s_j, \sigma_{-j}) = \mu(s_j, \tilde{\sigma}_{-j})$ for all $j \in I$ and $s_j \in S_j$ and $\mu(\tilde{s}_i, \tilde{\sigma}_{-i}) =$

$\mu(\sigma_i(\omega, t_i), \tilde{\sigma}_{-i}))$. Using the same argument as we did for BNE, it is immediate to argue $\tilde{\sigma}$ satisfies the inequality above for all $i \in I$, $t_i \in T_i$ and $s'_i \in \tilde{S}_i$. Strategy profile $\tilde{\sigma}$ is thus a cursed equilibrium of $\tilde{\gamma}$, concluding the proof.

Cursed Equilibrium satisfies IIA as well. Let σ be any cursed equilibrium of γ, and and $\tilde{\gamma} = (\tilde{\mu}, \tilde{S})$ with:

$$\tilde{S}_i = \{s_i \in S_i : \sigma_i(\omega, t_i)[s_i] > 0\}$$

For all $i \in I$ and $\tilde{\mu}(s) = \mu(s)$ for all $s \in S$. Consider now $\tilde{\sigma} \in \tilde{\Sigma}$ such that $\tilde{\sigma}_i(t_i)[s_i] = \sigma_i(\omega, t_i)[s_i]$ for all $i \in I$ and $s_i \in \tilde{S}_i$. It is immediate then to notice that, as $\sigma \in \mathcal{S}(\gamma)$ and $\tilde{\sigma}$ is payoff-equivalent to σ, that $\tilde{\sigma} \in \mathcal{S}(\tilde{\gamma})$. This completes the proof that the revelation principle holds for cursed equilibrium.

3.1.5 Interim Correlated Rationalizability

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:

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

By Tarski's theorem, there exists a largest fixed point of b which is denoted as $C^{\gamma(T)}$. We say σ is a solution to mechanism γ (or, equivalently, that it is rationalizable in γ) whenever $\sigma_i(\omega, t_i) \in b_i(C^{\gamma(T)})[t_i]$.

It is immediate to notice ICR satisfies separation invariance by noticing any action that is rationalizable in γ is also rationalizable in $\tilde{\gamma}$ by the same belief. Therefore, if $\sigma \in \mathcal{S}(\gamma)$, it follows $\tilde{\sigma}$ as defined above is a solution to $\tilde{\gamma}$ given that action $\tilde{s}_i$ in mechanism $\tilde{\gamma}$ yields the

same outcome as action $\sigma_i(\omega, t_i)$ in mechanism γ, regardless of the profile of action played by i's opponents.

ICR generally fails to satisfy IIA instead. As a matter of fact, type t_i of agent i may be able to rationalize action $\sigma_i(\omega, t_i)$ only through the belief their opponents will play a strategy different from $\sigma_{-i}(\omega, t_i)$. To see the point more clearly, consider the following example. For $i \in \{R, C\}$, let $T_i = \{-1, 1\}$ and assume each type profile occurs with the same probability. Let $A = \{a, b, c\}$, and let agent C derive utility t_C from every alternative. Agent R, instead, gets utility t_R from alternative a, 0 from alternative b, and 1 from alternative c.

Consider now the following mechanism, where $S_R = \{U, D\}$ and $S_C = \{L, M, R\}$:

	L	M	R
U	a	b	b
D	b	c	c

Which corresponds to the following payoff matrix:

	L	M	R
U	(t_R, t_C)	$(0, t_C)$	$(0, t_C)$
D	$(0, t_C)$	$(1, t_C)$	$(1, t_C)$

As C derives the same utility from all alternatives, all her actions are rationalizable for both her types. For player R of type $t_R = -1$, instead, the only rationalizable action is D. For type $t_R = 1$, U is rationalized by the belief C will play U and D by the belief her opponent will play R. In particular, in order for U to be rationalized, we need C to put at least probability $\frac{1}{2}$ on L.

Therefore, σ such that $\sigma_R(1)[U] = 1$, $\sigma_R(-1)[U] = 0$, $\sigma_C(1)[M] = 1$ and $\sigma_C(-1)[R] = 1$ is a solution to the mechanism. Notice this entails that the mechanism implements outcome b whenever $t_R = 1$ and c whenever $t_R = -1$.

However, it is not a solution to the reduced game:

	M	R
U	$(0, t_C)$	$(0, t_C)$
D	$(1, t_C)$	$(1, t_C)$

As a matter of fact, U is strictly dominated in the reduced game and is therefore not rationalizable for type $t_R = 1$. Therefore, σ is not a solution to this reduced game.

While a violation of IIA is not enough to conclude ICR does not satisfy the revelation principle, this counterexample provides us with a valuable starting point. As a matter of fact, we can obtain the direct mechanism associated with σ by relabeling the action space above to match with the type space:

	1	-1
1	b	b
-1	c	c

	1	-1
1	$(0, t_C)$	$(0, t_C)$
-1	$(1, t_C)$	$(1, t_C)$

Any solution of the direct mechanism will then yield c as an outcome, as in the direct mechanism -1 is a dominant strategy for R. Thus any solution will lead to outcome c regardless of the type profile, proving the SCF $\mu(\sigma)$ cannot be implemented via a direct mechanism. This concludes the proof.

The violation of IIA, in this case, stems from the fact agents assign a positive probability to the event an action that is not part of the solution will be played: as a result, removing that action from the corresponding agent's action profile implies no solution of the reduced mechanism presents the same distribution over actions as the original one.

3.2 Level-k Reasoning

In this subsection we will follow the discussion of level-k reasoning models presented in Kneeland (2022). Let us assume that Ω captures the strategic uncertainty the planner faces due to the fact agents may be of different levels:

$$\bar{\Omega} = \times_{i \in I} \bar{\Omega}_i = \times_{i \in I} \{k \in \mathbf{N} : k_i \leq \bar{K}\}$$

Adapting the definition in Kneeland (2022) to our setup, we can then say $\sigma = \times_i \sigma_i$ such that $\sigma_i : \bar{\Omega}_i \times T_i$ is a *level-k solution* to mechanism γ whenever for all $i \in I$, $t_i \in T_i$, $s_i' \in \Delta(S_i)$ and $\omega_i \in \bar{\Omega}_i / \{0\}$:

$$\int_{T_{-i}} u_i(\mu(\sigma_i(\omega_i, t_i), \sigma_{-i}(\omega_i - 1, t_{-i})), t)\, dp_i(t_{-i}|t_i) \geq \int_{T_{-i}} u_i(\mu(s_i', \sigma_{-i}(\omega_i - 1, t_{-i})), t)\, dp_i(t_{-i}|t_i)$$

That is, σ is a level-k solution whenever it prescribes agents of level k to best respond to the belief their opponents all are of level $k-1$.[11]

OC can be easily established by an argument similar to the ones presented above.

This formulation of level-k reasoning satisfies IIA as well: as a matter of fact, each level of each agent best responds to a mixed action that is played with non-zero probability in the solution. Therefore, no mixed actions other than the ones played in solution σ are relevant.

This result, however, relies on the assumption $\bar{\Omega}$ contains all non-negative profiles of levels bounded above by $\bar{K}$. To show why, generalize slightly the definition of a level-k solution to allow for $\Omega \subset \bar{\Omega}$ and assume σ is a level-k solution of mechanism γ for strategic state space Ω whenever there exists a level-k solution extending σ from Ω to $\bar{\Omega}$.[12]

To illustrate the point, suppose $\bar{K} = 1$, $\Omega = \{(1,1)\}$ and consider again the same setup and mechanism as in the previous section. Consider again the game in the previous section.

[11] In this setup, $\times_{i \in I} \sigma_i(0, t_i)$ acts as an anchor.

[12] Notice this definition coincides with the one of a level-k solution for $\Omega = \bar{\Omega}$

Notice one level-k solution for Ω is for the level-1 column player to pick R regardless of her type and for the level-1 row player to choose U when her type is $t_R = 1$ and D otherwise.[13] As the column player only plays R, we should be able to eliminate L without affecting our predictions on the play of the mechanism. As discussed above, this is not the case: IIA is violated.

Finally, we can check level-k reasoning satisfies SI. Constructing $\tilde{\gamma}$ and $\tilde{\sigma}$ as in the definition of SI, it is immediate to notice if σ is a level-k solution then $\tilde{\sigma}$ is as well. Applicability of the revelation principle to level-k reasoning models depends, however, on whether we allow the planner to (implicitly) pick the anchor for level-0 players. Take the anchor to be exogenous instead: for instance, suppose each agent's anchor is to randomize uniformly over her action set. In this case, duplicating an action might affect the best reply of level 1 players, entailing SI does not hold in general.

4 Necessity of Incentive Compatibility

The second aspect of the revelation principle this paper investigates is related to necessity of Bayesian Incentive Compatibility (BIC) for partial implementation. This aspect is particularly important as many key impossibility results in mechanism design are derived through the use of BIC: for instance, the impossibility result of Myerson and Satterthwaite (1983) or the Revenue Equivalence Theorem in Myerson (1981).

This section generalizes the set of results in Rubbini (2023) by extending them to partial implementation environments. We find similar conditions yield necessity of BIC for partial implementation and, by the means of a few examples, we provide intuition about settings in which such conditions may plausibly hold. The intuition behind the result presented in this section is similar to the one Rubbini (2023) discusses in the context of full implementation:

[13] To see why this is the case, it is enough to consider any extension of σ to $\bar{\Omega}$ in which the level-0 column player plays action L.

BIC is required for partial implementation when agents can expect the planner to select a particular solution to the implementation problem. In that case, agents' non-rational expectations about her opponents' *actions* are immaterial as long as their expectations about the *outcome of the mechanism* are correct.

In the context of a generalized principal-agents model, Myerson (1982) models selection of a solution (in that case, an equilibrium) by assuming the planner can recommend a strategy to each agent to coordinate them on the solution of interest. Even if the principal has no control over the action the agents take, they will follow the planner's recommendation as long as they are incentive compatible.[14]

In the same spirit, we assume the planner can select one of the solutions of the mechanism to implement the social choice function of interest by sending a message to the agents. To model this feature, we augment each mechanism γ with a pre-play communication stage in which the planner is allowed to issue a message $m \in M(\gamma)$ to all agents $i \in I$ at the same time she commits to the implementing mechanism γ. $M(\gamma)$ represents the set of all messages available to the planner when she commits to mechanism γ: for instance, it could be a fixed set $M = \times_{i \in I} M_i$ or $M(\gamma)$ could coincide with the set of γ's strategy profiles Σ. In the latter case, we can interpret the message as a recommendation to play according to a given strategy profile $\sigma \in \Sigma$, an intuition similar to Myerson (1982). By assuming $M(\gamma)$ is a singleton, we can interpret m as a social norm or focal solution too: for example, truthful reporting in a direct mechanism.

After learning their type and receiving the message, agents choose a strategy. Denote as $\tilde{\mathcal{S}}$ the solution concept for this extended mechanism, mapping each pair $(\gamma, m) \in (\Gamma, M(\gamma))$ into a subset of $\mathcal{S}(\gamma)$. This is consistent with the idea the planner is just able only to *select* a profile that was already a solution to the mechanism, but not to induce new solution profiles. The planner enjoys then an additional degree of freedom with respect to the standard full

[14] See Myerson (1982) for a discussion of the connection of this solution concept to correlated equilibrium Aumann (1974).

implementation model: in addition to committing to a mechanism, she can select a solution of this mechanism by communicating with the agents.

Similarly to how we defined a theory of expectations for a mechanism γ, we can define an *extended theory of expectations* $\tilde{E}(\gamma, m) \subseteq E(\gamma)$ for all $\gamma \in \Gamma$, so that the set of solutions to this extended mechanism is $\tilde{\mathcal{S}}(\gamma, m) = R(\tilde{E}(\gamma, m))$. At this moment, let us assume the planner's message affects only agents' expectations, but it does not affect how these expectations are mapped into strategies by the response correspondence. It is then immediate to notice that $\tilde{E}(\gamma, m) \subseteq E(\gamma)$ implies $\tilde{\mathcal{S}}(\gamma, m) \subseteq \mathcal{S}(\gamma)$.

Notice this framework captures the possibility of public and private messages by different formulations of $\tilde{\mathcal{S}}$. For example, we can model private messages by assuming $M(\gamma) = \times_{i \in I} M_i(\gamma)$ and that $\tilde{\mathcal{S}}(m, \gamma) = \times_{i \in I} \tilde{\mathcal{S}}_i(m_i, \gamma)$ for all $m \in M(\gamma)$ and $\gamma \in \Gamma$. Similarly,

We will say a SCF is *partially implementable* in $\mathcal{S}$ whenever there exists a pair (γ, m) such that $\mu(\tilde{\mathcal{S}}(\gamma, m)) = f$. We denote the set of all pairs (γ, m) implementing f as $\tilde{\Gamma}^f$. Similarly, we say a SCS F is *partially implementable* whenever there exists a pair (γ, m) such that $\mu(\tilde{\mathcal{S}}(\gamma, m)) = F$ and we denote the set of all such pairs as $\tilde{\Gamma}^F$.

We say $\tilde{\mathcal{S}}$ is *weakly consistent* (WC) for a mechanism-message pair (γ, m) whenever $\tilde{\mathcal{S}}(\gamma, m) \neq \emptyset$ and for all $i \in I$ and $t_i \in T_i$ there exists $\sigma \in \tilde{\mathcal{S}}(\gamma, m)$ such that:

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

We say $\tilde{\mathcal{S}}$ is weakly consistent whenever it is weakly consistent for all pairs (γ, m).

The interpretation of WC is similar to the one for Weak Response Consistency (WRC) in Rubbini (2023). That is, conditional on their expectations, agents believe they can induce a different outcome of the mechanism (among the ones implementing the social choice function chosen by the planner) by choosing the action a different type would pick. As for WRC, we can interpret WC as requiring some level of consistency between the model the planner has on how agents play the game (the solution concept) and the model the agent themselves

have. For a more detailed discussion of possible justifications for weak consistency, see the examples proposed in the next section.

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, BIC holds whenever no type of each agent has no incentive to pretend to be of a different type in the direct mechanism associated with the SCF.

The first result of this section establishes necessity of BIC for partial implementation when the solution concept is weakly consistent. This finding generalized the one obtained in Rubbini (2023) with similar arguments.

Theorem 3. *If f is implementable in $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{S}}$ is WC for $\tilde{\Gamma}^f$, then f is BIC. If a BIC f is implementable in $\tilde{\mathcal{S}}$, then $\tilde{\mathcal{S}}$ is WC for $\tilde{\Gamma}^f$.*

Theorem 1 also means there exist solution concepts that do not satisfy IIA but for which BIC is necessary for implementation. : that is, solution concepts for which BIC is necessary for implementation but for which there is a loss of generality in focusing on the class of direct mechanisms. These two aspects, which are conflated in the classical revelation principle, can then be separated and attributed to two different properties of the solution concept.

4.1 Examples

In this section, we will propose some examples of how agents' expectations may update after receiving a message from the planner.

4.1.1 Disregard for the message

The most extreme case to consider is when the planner's message does not affect agents' expectations at all. In this case, $E(\gamma, m) = E(\gamma)$ for all $\gamma \in \Gamma$ and $m \in M(\gamma)$. This implies only fully implementable SCFs are partially implementable, as $\tilde{\mathcal{S}}(\gamma, m) = \mathcal{S}(\gamma)$ for all γ and m. Moreover, $\tilde{\mathcal{S}}(\gamma, m)$ is weakly consistent for all $\gamma \in \Gamma$ for which $\mathcal{S}$ is Weakly Response Consistent (Rubbini, 2023).

4.1.2 BNE and Correlated Equilibrium

A popular argument to justify the use of partial as opposed to full implementation relies on the tenet the planner could suggest the agents which strategy to play and, as long as it is incentive compatible, the agents will follow the suggestion.[15]

This intuition can easily be captured in our model. Let us first characterize BNE by defining the following pair of theories of expectation and response (Rubbini, 2023):

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

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

We can now extend $\mathcal{S}^{BN}$ by letting $M(\gamma) = \Sigma$ and defining the extended theory of expectations as follows:

$$E^{BN}(\gamma, m) = \{e \in E^{BN}(\gamma) : e_i = m_{-i} \text{ for all } i \in I\}$$

This formulation captures the intuition that all agents expect their opponents to abide by the message sent by the planner: as argued in Myerson (1982), this constitutes a form of correlated equilibrium. Notice that $\tilde{\mathcal{S}}^{BN}$ is weakly consistent for any γ and m such that

[15] In this case, the planner induces a correlated equilibrium of the mechanism augmented by the pre-play communication stage. See Myerson (1982) for a complete discussion.

$m \in \mathcal{S}^{BN}(\gamma)$ as BNE is a WRC solution concept.

4.1.3 Planner's Utility Maximization

Weak Consistency can also be justified by assuming agents can accurately predict what SCF the planner seeks to implement.[16] For instance, the planner may design a mechanism to implement a law-mandated SCF, or the SCF may be the only allocation rule that maximizes an auction's revenue.

In order to capture this intuition, assume the set $\mathcal{F}$ of all functions $f : T \to \Delta(A)$ is endowed with a linear order. We can interpret this linear order as a preference relation on the set of social choice functions. To keep matters simple, let me assume every subset of $\mathcal{F}$ admits a maximal element in this order. Under this simplifying assumption, planner's preferences can be represented by an utility function $v : \mathcal{F} \to \mathbb{R}$ that admits a non-empy set of (local) maxima on every subdomain $F \subseteq \mathcal{F}$. For every $F \subseteq \mathcal{F}$, let $F_v^* = \arg\max_{f \in F} v(f)$ denote the set of such maxima and, for each mechanism γ, denote as $\mathcal{S}_v^*(\gamma) = \{\sigma \in \Sigma : \mu(\sigma) \in F_v^*\}$ the set of solutions implementing one of the planner's preferred SCFs. This setup can also be interpreted as the planner being mandated by law to implement such a SCF, but being able to choose the implementing mechanism freely. For example, if f is the SCF the planner mandated to implement, we could set $v(f) = 1$ and $v(f) = 0$ otherwise.

For the remainder of the section, we will assume the planner's choice of a mechanism and a message aims at maximizing her utility v and that this is known to the agents. Let the solution concept for the extended mechanism be denoted with $\tilde{\mathcal{S}}^* = R(E^*)$, where E^* denotes the subset of expectations that are consistent with agents believing that, by pretending to be of a different type, they can induce a different solution among those that

[16] This intuition is exactly the same as the one underlying of the results about full implementation in Rubbini (2023).

maximize the planner's utility. That is, for a given E, $R \subseteq B$ and $m \in M$:

$$E^*(\gamma, m) \subseteq \{e \in E(\gamma) : (R_{i,t_i}(e_{i,t_i}), e_{i,t_i}) \in \mathcal{S}_v^*(\gamma) \text{ for all } i \in I, t_i \in T_i\}$$

Notice imposing this restriction on $E^*(\gamma, m)$ entails $\tilde{\mathcal{S}}^*(\gamma, m) \subseteq \mathcal{S}_v^*(\gamma)$ for all pairs (γ, m): that is, agents know that replying to their expectations will lead to one of the outcomes of the mechanism that maximizes the planner's utility (that is, a SCF in F_v^*). In other words, agents believe the mechanism γ the planner commits to and the message m she sends are chosen to maximize her utility v. This presumes agents know what the objectives of the planner are, i.e. that the planner does not possess any private information about what her payoffs are. In particular, if every agent knows v and v admits a unique maximizer, all agents would believe the pair (γ, m) implements the same SCF.

Theorem 4. *If f is implementable in $\tilde{\mathcal{S}}^*$ via (γ, m) and v admits a unique maximizer in the set $F = \mu(\mathcal{S}(\gamma))$, $\tilde{\mathcal{S}}^*$ is WC for $(\gamma, M(\gamma))$.*

4.2 Applications

The results of previous sections are of particular importance as the insights behind BIC (and related properties) turn out to be crucial in many setups: we can then exploit the results from the previous section to appreciate how some classical results extend to partial implementation in a solution concept that does not satisfy rational expectations. For the purpose of the discussion of this section, let me assume the setup is the same as in the utility maximization framework described above.

4.2.1 Bilateral Trading

This section extends and complements the analysis of the impossibility results of Myerson and Satterthwaite (1983), already explored in de Clippel et al. (2019), Crawford (2021) and

Rubbini (2023).

As in the paper from Myerson and Satterthwaite (1983), 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. This value t_i is distributed according to some distribution P_i, which we assume admits a continuous and positive pdf over the interval $[a_i, b_i]$. I also assume that t_B is independent of t_S and that each agent knows her valuation and does not know the one of the other agent but knows its distribution. Following Myerson and Satterthwaite (1983), I assume $(a_S, b_S) \cap (a_B, b_B) \neq \emptyset$ as well.

The set of alternatives consists of all pairs (q, x), where $q \in [0, 1]$ represents the probability 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.

We say a SCS is *individually rational* if for all $f \in F$, $i \in I$ and $t \in T$:

$$u_i(f(t), t) = q(t)t_i - x(t) \geq 0$$

We say a SCS is *ex-post efficient* whenever f is such that $q(t) = 1$ whenever $t_B > t_S$ and $q(t) = 0$ whenever $t_B < t_S$. Under these conditions, Myerson and Satterthwaite (1983) show that there exists no SCF f that is simultaneously individually rational, ex-post efficient, and incentive compatible.

As their result relies on invoking the revelation principle and studying only the direct mechanism (f, T), it follows any SCF f maximizing v for the planner is not implementable.

Corollary 2. *Let f be individually rational and ex-post efficient. If $\mu(\mathcal{S}_v^*(\gamma))$ is a singleton for all $\gamma \in \Gamma$, then f is not implementable in $\mathcal{S}^*$.*

When would $\mu(\mathcal{S}_v^*(\gamma))$ plausibly be a singleton? As an example, we can us assume the

planner's utility is given by:

$$v(f) = \int_{t \in T} w(u_B(f(t), t), u_S(f(t)), t))\, dp(t)$$

where $w : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is strictly quasi-concave for all $t \in T$: that is, in each state, the planner prefers less extreme distribution of surplus to more extreme ones.[17]

4.2.2 Revenue Maximizing Auctions

We can use the planner's utility maximization framework above to discuss revenue maximization in single-unit auctions. We show that if agents know the planner is maximizing utility under some pretty natural constraints (such as individual rationality and anonymity), then it is not possible to implement the any SCF that maximizes revenue in the class of individually rational social choice functions.

Define the set of alternatives 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 of winning the object q_i and a monetary transfer x_i. 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.

Agent's Bernoulli utilities for each agent $i \in I$ and type $t_i \in T_i$ are:

$$u_i((q_i, x_i), t) = q_i v_i(t) - x_i(t)$$

Where each agent's valuation v_i is increasing in t_i and non-negative. Assume moreover agents' types are distributed according to a commonly known distribution p with support

[17] Alternatively, we could assume the planner prefers more extreme distributions of surplus: for instance, when the planner prefers the buyer (or the seller) to reap all the surplus from trade.

$[\ \underline{t},\ \bar{t}\ ]^{|I|}$, where $\bar{t} > \underline{t} \geq 0$.

Let us say a SCF f is *anonymous* whenever for all permutations $\pi : I \to I$, $q_i(t_1, ..., t_I) = q_{\pi(i)}(t_{\pi(1)}, ..., t_{\pi(I)})$: that is, the probability of i winning the object does not depend on her identity, but only on the her and her opponents' types.

We will then assume the planner wants to implement the SCF f maximizing her expected revenue in the class of individually rational and anonymous SCFs. That is, the planner then seeks to implement f such that:

$$\begin{aligned} \max_{f} \quad & \sum_{i \in I} \left(\int_T x_i^f(t)\, dp(t) \right) \\ \text{s.t.} \quad & \int_{T_{-i}} (q_i^f(t) v_i(t) - x_i^f(t))\, dp_i(t_{-i}|t_i) \geq 0 \text{ for all } i \in I, t_i \in T_i \\ & f \text{ is anonymous} \end{aligned}$$

Suppose this goal of the planner is known to the agents: we can model such a feature by assuming $\mathcal{S}_v^*(\gamma)$ coincides with the solutions of γ that are individually rational, anonymous, and maximize revenue in the mechanism.

Finally, let us rule out the uninteresting case in which the planner could maximize revenue by just setting up a posted-price mechanism with price equal to $\max_{i \in I} v_i(\bar{t})$. That is, we assume the SCF f solving the maximization problem above is satisfies:

$$\sum_{i \in I} \left(\int_{T_{-i}} x_i^f(t)\, dp(t) \right) > \max_{i \in I} p(\{t \in T : t = \bar{t}\}) v_i(\bar{t})$$

In this case, we show the planner is unable to implement the SCF f that maximizes revenue subject to the IR and anonimity constraints.

Corollary 3. *If a SCF f is revenue-maximizing in the class of IR and anonymous SCFs, then it is not implementable in $\tilde{\mathcal{S}}^*$.*

Proofs

Proof of Theorem 1. Let $\sigma \in \mathcal{S}(\gamma)$. We will now use IIA and SI to prove there exists a bijection from T to S that is outcome-equivalent to σ.

We construct mechanism $\gamma' = (\mu', S')$ as follow. Let $S' = S \cup \{\Omega \times T_i\}$ and, for all $K \subseteq I$, $t_K \in T_K$ and $s_{-K} \in S_{-K}$, let $\mu'((\omega, t_i), s_{-i}) = \mu(\sigma_i(\omega, t_i), s_{-i})$. Notice any action in $\{\Omega \times T_i\} = S'/S$ is a duplicate of mixed action $\sigma_i(\omega, t_i) \in \Delta(S)$, so that γ' is a redundant extension of γ. Then there exists σ' such that $\sigma'_i(\omega, t) \neq \sigma'_i(\omega, (t'_i, t_{-i}))$ and $\sigma'_i(\omega, t) \in D(\sigma_i(\omega, t))$ for all $i \in I$, $t_i, t'_i \in T_i$ and $t_{-i} \in T_{-i}$.

By SI, there exists $\sigma' \in \mathcal{S}(\gamma')$ that is injective and such that $\mu(\sigma) = \mu'(\sigma')$. Moreover, we can reduce the strategy space of γ' to the support of σ' by constructing $\gamma'' = (\mu', \sigma'(\Omega, T))$. By IIA, there exists $\sigma'' \in \mathcal{S}(\gamma'')$ such that $\sigma''(\omega, t)[\sigma'(\omega, t)] = 1$. Notice moreover that $\mu''(\sigma'') = \mu'(\sigma') = \mu(\sigma)$.

Consider now $\gamma^D = (\mu \circ \sigma, T)$. Let $\rho : S'' \to T$ be the bijection induced by σ'' that associates each element of S'' with the corresponding element of T. Notice also $f(\rho) = \mu''$. Therefore by OC we have:

$$\sigma'' \in \mathcal{S}(\gamma'') \implies \rho(\sigma'') \in \mathcal{S}(\gamma^D)$$

This implies that the identity function is a solution of γ^D, proving truth-telling is a solution of the direct revelation mechanism. We then conclude the proof. □

Proof of Theorem 2. Necessity of BIC for implementation follows from the usual argument. As for direct monotonicity, suppose f is fully implementable in pure BNE via a direct mechanism. Then there exists $\sigma \in \mathcal{S}((f, T))$ such that $f(\sigma) = f$. As there exists no other pure BNE with a different outcome, it must be that for any $\alpha : T \to T$ with $f \neq f(\alpha)$,

there exist $t_i' \in T_i$ and $i \in I$ such that:

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

Moreover, as σ is a BNE, we have that for all $i \in I$ and $t_i, t_i' \in T_i$:

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

Which implies:

$$\int_{T_{-i}} u_i(f(t), t)\, dp_i(t_{-i}|t_i) \geq \int_{T_{-i}} u_i(f(\alpha_i(t_i), t_{-i}), t)\, dp_i(t_{-i}|t_i)$$

Let us now prove the converse statement. BIC ensures truth-telling is a BNE of the direct mechanism. Moreover, direct monotonicity entails any deception $\alpha : T \to T$ such that $f \neq f(\alpha)$ cannot be an equilibrium of (f, T): therefore, all equilibria of the direct mechanism implement f. This concludes the proof. □

Proof of Theorem 3. As f is partially implementable, there exists γ and $m \in \Sigma$ such that $\mu(\tilde{\mathcal{S}}(\gamma, m)) = f$. By consistency, for all $i \in I$ and $t_i \in T_i$ there exists $e \in E(\gamma, m)$ and $\sigma_i \in \Sigma_i$ such that $\sigma_i(\omega, t_i) \in R_{i,t_i}(e_{i,t_i})$ and $\mu(\sigma_i, e_{i,t_i}) \in \mu(\mathcal{S}(\gamma, m))$.

Then, for all $i \in I$ and $t_i, t_i' \in T_i$, $\sigma_i(\omega, t_i) \in R_{i,t_i}(e_{i,t_i})$ implies:

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

By consistency and implementability, we have $f = \mu \circ (\sigma_i, e_{i,t_i})$ for all $i \in I$ and $t_i \in T_i$, so the inequality above can be rewritten:

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

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

Proof of Theorem 4. As (γ, m) implements f, $\tilde{\mathcal{S}}^*(\gamma, m) \neq \emptyset$ and therefore $E^*(\gamma, m) \neq \emptyset$. Moreover, for all $e \in E^*(\gamma, m)$ we have that for all $i \in I$ and $t_i \in T_i$ there exists $\sigma \in \Sigma$ with $\sigma_i(\omega, t_i) \in R_{i,t_i}(e_{i,t_i})$ and $(\sigma_i, e_{i,t_i}) \in S_v^*(\gamma)$. It follows that $\tilde{\mathcal{S}}^*(\gamma, m) \subseteq \mathcal{S}_v^*(\gamma)$ and then $\mu(\tilde{\mathcal{S}}^*(\gamma, m)) \subseteq \mu(\mathcal{S}_v^*(\gamma))$. As v has a unique maximizer, $\mu(\mathcal{S}_v^*(\gamma))$ is a singleton and $\mu(\tilde{\mathcal{S}}^*(\gamma, m)) = \mu(\mathcal{S}_v^*(\gamma))$. Weak consistency then follows from the fact that for all $i \in I$, $t_i \in T_i$ and $\omega \in \Omega$, $\sigma_i(\omega, t_i)$ is a best reply to $e_{i,t_i}(\omega, t_{-i})$. □

Proof of Corollary 2. Suppose f is implementable via the mechanism-message pair (γ, m). As $\mathcal{S}_v^*(\gamma)$ is a singleton for all $\gamma \in \Gamma$, $\mathcal{S}^*$ is WC and, by Theorem 4, f is BIC. As Myerson and Satterthwaite (1983) prove there exists no SCF that is simultaneously BIC, individually rational, and ex-post efficient, we conclude the proof. □

Proof of Corollary 3. Suppose for the sake of contradiction that f is implementable in $\tilde{\mathcal{S}}^*$ via mechanism-message pair (γ, m). As f maximizes revenue in the class of individually rational social choice functions, it must be that each agent's type has expected utility equal to her reservation level, i.e. for all $i \in I$ and $t_i \in T_i$:

$$\bar{x}_i^f(t_i) = \int_{T_{-i}} q_i^f(t) v_i(t)\, dp_i(t_{-i}|t_i) = \bar{z}_i^f(t_i)$$

If that was not the case, any $\tilde{f}$ such that $\bar{x}_i^{\tilde{f}}(t_i) = \bar{z}_i^{\tilde{f}}(t_i)$ would increase revenue for the planner and it would still be individually rational. Therefore, f maximizes revenue in the class of IR and anonymous auctions only if it extracts all surplus from trade. That is, only if the planner's revenue is equal to:

$$\sum_{i \in I} \left(\int_{T_{-i}} q_i^f(t) v_i(t)\, dp(t) \right) > \max_{i \in I} v_i(\bar{t}) \int_T$$

Moreover, if f is implementable, there exists $\sigma \in \tilde{\mathcal{S}}(\gamma, m)$ such that $\mu(\sigma) = f$. By definition

of $\tilde{\mathcal{S}}^*$ we then have, for all $i \in I$ and $t_i, t_i' \in T_i$:

$$\int_{T_{-i}} u_i(\mu(\sigma_i(t_i), e_{i,t_i}(t_{-i})), t_i)\, dp_i(t_{-i}|t_i) \geq \int_{T_{-i}} u_i(\mu(\sigma_i(t_i'), e_{i,t_i}(t_{-i})), t_i)\, dp_i(t_{-i}|t_i)$$

Let $\sigma' = (\sigma_i, e_{i,t_i})$. We can then rewrite the inequality above as:

$$\int_{T_{-i}} q_i^{\mu(\sigma')}(t_i) v_i(t)\, dp_i(t_{-i}|t_i) - \bar{x}_i^{\mu(\sigma')}(t_i) \geq \int_{T_{-i}} q_i^{\mu(\sigma')}(t_i') v_i(t)\, dp_i(t_{-i}|t_i) - \bar{x}_i^{\mu(\sigma')}(t_i')$$

As $\sigma' \in \mathcal{S}_v^*(\gamma)$, $\mu(\sigma')$ must generate at least as much revenue as $\mu(\sigma) = f$. As the payoff of any type of any player can not fall below 0 by individual rationality, it must be $\bar{x}_i^{\mu(\sigma')} = \bar{z}_i^{\mu(\sigma')}$ for all $i \in I$ and $t_i \in T_i$. It follows then from the inequality above that:

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

As v_i is strictly increasing in i's type, this entails $q_i^{\mu(\sigma')}(t_i', t_{-i}) = 0$ for all $t_{-i} \in T_{-i}$. By considering $t_i = \bar{t}$, this result implies $q_i^{\mu(\sigma')}(t) = 0$ for all $t_{-i} \in T_{-i}$ and $t_i' < \bar{t}$.

By anonymity of $\mu(\sigma')$, the same holds true for all $j \neq i$ as well. Therefore, the revenue from the SCF f maximizing revenue in the class of IR and anonymous SCFs is:

$$\sum_{i \in I} \left(\int_{T_{-i}} x_i^f(t)\, dp(t) \right) > \max_{i \in I} p(\bar{t}) v_i(\bar{t})$$

This contradicts our premise and concludes the proof. □

References

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

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

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

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

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

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

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

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

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

Myerson, R. B. (1982). Optimal coordination mechanisms in generalized principal–agent problems. *Journal of Mathematical Economics 10*(1), 67–81.

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

Palfrey, T. R. and S. Srivastava (1989). Implementation with incomplete information in exchange economies. *Econometrica 57*(1), 115–134.

Postlewaite, A. and D. Schmeidler (1986). Implementation in differential information economies. *Journal of Economic Theory 39*(1), 14–33.

Rubbini, G. (2023). Mechanism design without rational expectations. *Bravo Center Graduate Student Working Paper Series 2023-001*.

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

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