Level-$k$ Mechanism Design

Geoffroy de Clippel  Rene Saran
Brown University  Yale-NUS College
Roberto Serrano
Brown University

This version: December 2016

Abstract
Models of choice where agents see others as less sophisticated than themselves have significantly different, sometimes more accurate, predictions in games than does Nash equilibrium. When it comes to the maximal set of functions that are implementable in mechanism design, however, they turn out to have surprisingly similar implications. Focusing on single-valued rules, we discuss the role and implications of different behavioral anchors (arbitrary level-0 play), and prove a level-$k$ revelation principle. If a function is level-$k$ implementable given any level-0 play, it must obey a slight strengthening of standard incentive constraints (we term it strict-if-responsive Bayesian incentive compatibility (SIRBIC)). Further, the same condition is also sufficient for level-$k$ implementability, although the role of specific level-0 anchors is more controversial for the sufficiency argument. Nonetheless, our results provide tight characterizations of level-$k$ implementable functions in independent private values and general environments, under a variety of level-0 play, including truthful, uniform, and atomless anchors.

JEL Classification: C72, D70, D78, D82.
Keywords: mechanism design; bounded rationality; level $k$ reasoning; revelation principle; incentive compatibility.

*We thank Sandro Brusco, Antonio Cabrales, Claude d’Aspremont, Navin Kartik, Ludovic Renou, and Yuval Salant for comments and suggestions.
1 Introduction

Models of choice where agents see others as less sophisticated than themselves have significantly different, sometimes more accurate, predictions in games than does Nash equilibrium. Evidence suggests that theories of level-$k$ choice may provide a better description of people’s behavior, especially when they are inexperienced.\footnote{See, for example, Stahl and Wilson (1994, 1995), Nagel (1995), Ho et al. (1998), Costa-Gomes et al. (2001), Bosch-Domènech et al. (2002), and Arad and Rubinstein (2012).} There are two main points that this paper makes: (i) when it comes to identifying the maximal set of functions that can be implementable in mechanism design, the two approaches turn out to have surprisingly similar implications, and (ii) to determine whether a given function is implementable requires to take a stand with respect to additional assumptions regarding what should come first, the implementing mechanism or the assumed level-0 play therein.

Mechanism design aims at engineering rules of interaction that guarantee desired outcomes while recognizing that participants may try to use their private information to game the system to their advantage. The design problem thus hinges upon a theory of how people make choices given the rules that are being enforced. Oftentimes the concept of Nash equilibrium is used for that purpose, but the past few years have seen a number of papers incorporating lessons from behavioral economics into mechanism design.\footnote{For instance, Eliaz (2002) allows for “faulty” agents, Cabrales and Serrano (2011) allow agents to learn in the direction of better replies, Saran (2011) studies the revelation principle under conditions over individual choice correspondences over Savage acts, Renou and Schlag (2011) consider implementation with $\epsilon$-minmax regret to model individuals who have doubts about others’ rationality, Glazer and Rubinstein (2012) allow the content and framing of the mechanism to play a role, de Clippel (2014) relaxes preference maximization, and Saran (2016) studies $k$ levels of rationality with complete information.}

The Nash equilibrium and level-$k$ approaches assume that participants are rational, to the extent that they maximize their preferences given their beliefs regarding how others will play. The difference lies in how beliefs are determined. Level-$k$ theories break down the Nash-equilibrium rational-expectations logic by assuming people see others as being less sophisticated than themselves. Best responses then determine behavior by induction on
the individuals’ depth of reasoning, starting with an “anchor” that fixes the behavior at level-0. This anchor captures how people would play the game instinctively, as a gut reaction without resorting to rational deliberation. We shall be careful in considering a host of possibilities for such behavioral anchors.

The revelation principle (see, e.g., Myerson (1989) and the references therein) offers an elegant characterization of the social choice functions that are (weakly) Nash implementable. Indeed, there exists a mechanism with a Bayesian Nash equilibrium that generates the social choice function if and only if the function is Bayesian incentive compatible, which means that telling the truth forms a Bayesian Nash equilibrium of the corresponding direct revelation game. How does level-\( k \)-implementation compare to this benchmark? To tackle this question, we need to be more precise regarding what we mean by level-\( k \) behavior. For ease in the exposition, we choose to concentrate on the level-\( k \)-reasoning model, in which each level-\( k \) individual best-responds to her belief that all her opponents are of level-\( (k - 1) \). We will then observe, in the concluding remarks section, that our results apply to numerous other specifications of bounded levels of reasoning.

In any theory that relies on bounded levels of reasoning, predictions depend on how one sets the level-0 anchor. Whether the mechanism designer can impact this anchor is debatable. Implementation is most permissive when giving her the freedom to pick the anchor. Thus, it should come as a surprise at first that, even with that power, the mechanism designer can implement only Bayesian incentive compatible social choice functions under level-\( k \) reasoning. This is our main result (Theorem 1), which amounts to a level-\( k \) revelation principle. In fact, the restriction is slightly stronger, as the incentive constrains must be satisfied with an inequality whenever the social choice function is responsive. We term this condition \textit{strict-if-responsive Bayesian incentive compatibility} (SIRBIC). Theorem 1 thus is the surprising assertion that if a social choice function is implementable up to level-\( K \) for \( K \geq 2 \) and for \textit{any} behavioral anchor at level-0, then it must satisfy SIRBIC. Our notion of implementability excludes outcomes produced by level-0 play (following Craw-
ford (2014), level-0 captures the beliefs that exist only in the minds of level-1’s and upper levels; these beliefs are about other players’ gut reactions to the game).

For an intuition behind Theorem 1, note that each player’s level-$k$ strategy is a best response to other players’ level-$(k - 1)$ strategies, but whenever $k > 1$, this level-$k$ strategy composed with those level-$(k - 1)$ strategies must implement the social choice function, that is, due to the definition of implementation, the resulting outcome is the same as the truth-telling outcome under the social choice function. Thus, although players’ beliefs about other players’ strategies are not consistent, as they are in equilibrium, for $k > 1$, level-$k$ of each player consistently believes that it can at best get the truth-telling outcome under the social choice function. Since the truth-telling outcome is the best for him, then, in particular, lying when others tell the truth cannot be better. This is exactly Bayesian incentive compatibility (an additional step is required to take us to SIRBIC).

One can generate, as in Crawford’s (2016) treatment of the Myerson and Satterthwaite (1983) bilateral trading mechanism, interesting examples of social choice functions that are implementable by level-$K$ play and fail to be Bayesian incentive compatible. One might thus be tempted to think that bounded levels of reasoning could free us from standard incentive constraints, but it follows from our level-$K$ revelation principle that this is in general a false hope. Indeed, such examples must be very special: it must necessarily be true that either $K = 1$ or that lower levels $k < K$ are excluded when evaluating the outcomes; see Example 1. Nonetheless, and remarkably, under suitable assumptions on level-0 and on type distributions (that is, the type distribution coincides with the distribution over messages generated by the level-0 anchor in the direct mechanism), Crawford (2016) manages to implement with level-$K$ play outcomes that are close to efficiency, establishing that bounded levels of reasoning perform close to equilibrium in this important sense. Gorelkina (2015) offers similar results in the d’Aspremont and Gerard-Varet’s (1979) expected externality mechanism.

After obtaining the level-$k$ revelation principle for arbitrary anchors, which
identifies the functions that satisfy SIRBIC as the only ones that can be level-\(k\) implementable, the rest of the paper makes concrete assumptions on behavioral anchors. This step is not free of controversy. Indeed, if one is interested in implementing a specific function, one can use the function itself as a direct mechanism, or design a different mechanism that will do the job. But then, it is possible that the level-0 anchor will be mechanism-specific, and hence, it is unclear whether one should assume first that the mechanism is in place and study plausible level-0 anchors for it, or rather, one should get started with a given level-0 behavior, provided there is strong empirical support for it, and apply that to the mechanism in question. The literature so far has followed the latter route (e.g., Crawford (2016), Gorelkina (2015)), and we will continue to explore here down this same path, but this is a methodological point on which new contributions to this literature may well have to pause and rethink methods. Other difficulties may arise, including level-0 being player-specific or even type-specific.

With these caveats, the converse of Theorem 1 holds, as shown in our next result (Theorem 2). Thus, any social choice function satisfying SIRBIC can be level-\(k\) implementable, asserting that the restriction imposed by the bounded levels of reasoning does not result in the planner not being able to implement some of the functions in the maximal set identified by Theorem 1. The applicability of this converse, Theorem 2, would be limited if the anchors needed to achieve implementability were unreasonable. However, Theorem 2 is proved with truth-telling as an anchor in direct mechanisms, often invoked as focal, as argued below. We also report a variant of Theorem 2 if anchors in direct mechanisms are nontruthful.

Insisting on the caveat outlined above, we move beyond direct mechanisms. It is generally appropriate to dwell on the importance of using a variety of behavioral anchors for level-0. For instance, the literature has discussed the use of uniform behavioral anchors, in which the gut reaction to a mechanism is to play an action chosen uniformly at random from the available actions (Crawford (2016)), or anchors that match type distributional assumptions (Gorelkina (2015)). In our quest to draw conclusions that hold for a wide class of
behavioral anchors, we offer general results for any atomless anchors in mechanisms with a continuum of actions, which include a number of previously used level-0’s as particular cases.

Specifically, with independent private values, and without using direct mechanisms or truthful anchors, one can still level-k implement all continuous SIRBIC functions. That is, using atomless anchors in a mechanism we construct, SIRBIC alone suffices for level-k implementation of continuous social choice functions (Theorem 3).³ Thus, effectively, with independent private values, Theorems 1 and 3 together provide a characterization of the continuous social choice functions that are level-k implementable with atomless anchors, and the key condition continues to be SIRBIC.

To obtain a similar characterization beyond independent private values is somewhat more challenging. An additional weak necessary condition is uncovered for uniform or type-independent anchors, which amounts to the social choice reacting to types having different interim preferences (Theorem 4). Conversely, this measurability condition and SIRBIC are also sufficient if anchors are atomless (Theorem 5). Thus, Theorems 1, 4, and 5 together provide a characterization of level-k implementable functions under atomless type-independent anchors: only those SIRBIC rules that pass the weak measurability restriction are implementable (see Example 3).

To sum up, the take-away message of our work is two-fold. On one hand, incentive compatibility arises as the robust condition that is key to describe the scope of implementable social choice functions, even if bounded levels of reasoning are factored into the model. On the other hand, to determine whether a particular function is implementable under level-k reasoning, one must scrutinize the plausibility of the specific level-0 anchor assumed in the implementing mechanism. While we provide tight characterization results for level-k implementable functions, we are aware of the limitations of the approach followed thus far. We hope nevertheless that the landscape depicted here will be of interest to behavioral economists and mechanism design theorists, with the aim of making progress in the design of institutions used by agents who are

³Continuity can be dispensed with, as we discuss later.
less than fully rational.

In closing, we stress that there are several modeling choices that are behind our results:

(i) We are restricting attention to single-valued rules (see Example 2 for a departure from this assumption), in particular implying that each level of reasoning beyond level-0 is treated the same by the planner. One could relax this and allow for more flexibility, but then one would have to defend such a differential treatment, questionable on normative grounds because levels of reasoning do not determine individual preferences.

(ii) We take an agnostic approach, and assume that all levels of reasoning are possible under some arbitrary upper bound $K \geq 2$. Different results would emerge should the planner believe that individuals’ levels are not sufficiently heterogeneous (see Example 1).

(iii) Our results use a notion of full implementation (all behavior compatible with best responses to the previous level must agree with the social goal). We observe, though, that this follows the behavioral anchor assumed for level-0, which makes our results for a specific anchor resemble more those of partial implementation. Moreover, it is not clear what justification the designer could offer to limit his attention to only a subset of behavior compatible with each level. In standard models of partial implementation, equilibrium offers such a justification (e.g., truthful equilibrium in the direct mechanism is focal); similar considerations need to be better understood in level-$k$ theory.

In further work, it will be important to explore settings where some of these assumptions are removed.

The paper is organized as follows. Section 2 presents the framework. Section 3 defines level-$k$ implementation. Section 4 presents our general necessity result – the “level-$k$ revelation principle” and two important examples to stress its power and limitations. Section 5 presents our sufficiency results for truthful anchors used in direct mechanisms. Section 6 contains our treatment of uniform/atomless anchors, and Section 7 closes with several concluding remarks.
2 Framework

A social planner/mechanism designer wishes to select an alternative from a set $X$. Her decision impacts the satisfaction of individuals in a finite set $I$. Unfortunately, she does not know their preferences nor does she know their level of cognitive sophistication. We discuss the more standard aspects of the framework in the current section, and postpone our treatment of bounded rationality, central to our work, to the next section.

In order to capture general problems of incomplete information, for each individual $i$, we introduce a set $T_i$ of types, with the interpretation that each individual knows his own type, but not the types of others. Beliefs are determined by Bayes’ rule using a common prior $p$ defined over $T = \prod_{i \in I} T_i$. Thus, when individual $i$’s type is $t_i$, her belief regarding other individuals’ types is given by the conditional distribution $p(\cdot | t_i)$. An individual $i$’s preference is of the expected utility form, using a Bernoulli utility function $u_i : X \times T \to \mathbb{R}$.

With a slight abuse of notation, we will write $u_i(\ell, t)$ to denote the expected utility of a lottery $\ell \in \Delta X$, where $\Delta X$ is the set of probability distributions over $X$.

The planner’s objective is to implement a social choice function $f : T \to \Delta X$. To achieve this goal, she constructs a mechanism, which is a function $\mu : M_1 \times \cdots \times M_I \to \Delta X$, where $M_i$ is the set of messages available to individual $i$. A mechanism is direct if $M_i = T_i$, for all $i$. A strategy of individual $i$ is a function $\sigma_i : T_i \to \Delta M_i$, where $\Delta M_i$ is the set of probability distributions over $M_i$. A strategy profile $\sigma$ and type profile $t$ induce a lottery $\mu(\sigma(t))$ over $X$.

We make several technical observations. Throughout the paper, it is assumed that the sets and functions considered have the right structure to make sure that expected utility is well-defined. Formally, the set of alternatives, and the sets of types and messages for each individual are separable metrizable.

---

4We reserve the term “type” to describe an individual’s beliefs about the payoff state, as well as beliefs about such beliefs. Since an individual’s depth of reasoning impacts her behavior but not her preferences, we do not include the depth of reasoning in the description of types. This is without loss of generality, as we later account for how individual behavior varies with her depth of reasoning.

5Formally, for any Borel subset $B$ of $X$, $\mu(\sigma(t))[B] = \int_m \mu(m)[B]d\sigma(t)$. 

8
able spaces endowed with the Borel sigma algebra, product sets are endowed
with the product topology, the Bernoulli utility functions are continuous and
bounded, and social choice functions, mechanisms, and strategies are measurable
functions.

3 Level-k Implementation

Together with types, beliefs, and utility functions, a mechanism \( \mu \) defines a
Bayesian game. To discuss implementation, we need to introduce a theory of
how people play Bayesian games. We present our results in this section for
the level-k model. In the concluding section, we comment on how our results
can be extended to other alternative models of choice with bounded depth of
reasoning.

In order to describe choices, we begin by introducing behavioral anchors,
which describe how a given individual would instinctively play the mechanism,
as a gut reaction without any rational deliberation. Formally, individual \( i \)'s
behavioral anchor \( \alpha_i \) is a strategy that associates to each type \( t_i \) a probability
distribution over \( M_i \), i.e., a mapping \( \alpha_i : T_i \to \Delta M_i \), which, therefore, is
mechanism-contingent. Profiles of such anchors will be denoted \( \alpha = (\alpha_i)_{i \in I} \).
We remark that, at this point, the behavioral anchors are completely arbitrary,
and they may differ across agents.

The set of strategies that are level-1 consistent for an individual is then
the set of her best responses against the other individuals’ behavioral an-
chors, that is, \( S^1_i(\mu|\alpha) \) is the set of strategies \( \sigma_i \) such that \( \sigma_i(t_i) \) maximizes
\[
\int_{t \in T_i} u_i(\mu(m_i, \alpha_{-i}(t_{-i})), t)dp(t_{-i}|t_i) \quad \text{over } m_i \in M_i.
\]
By induction, for each \( k \geq 1 \), the set of strategies that are level-(\( k + 1 \)) consistent for an individual is the
set of her best responses against a strategy profile that is level-\( k \) consistent for
the other individuals, that is, \( S^{k+1}_i(\mu|\alpha) \) is the set of strategies \( \sigma_i \) such that
\( \sigma_i(t_i) \) maximizes \( \int_{t \in T_i} u_i(\mu(m_i, \sigma_{-i}(t_{-i})), t)dp(t_{-i}|t_i) \), for some \( \sigma_{-i} \in S^k_{-i}(\mu|\alpha) \).
The index \( k \) is called an individual’s depth of reasoning.

It has been argued that, for many subjects in the lab, their depth of rea-
soning is probably rather small. At the same time, this depth varies from
individual to individual, and even within a person, it may vary from mechanism to mechanism. It is currently not well understood how one could identify or impact individuals’ depth of reasoning. To accommodate this, we introduce an upper bound $K$ on the individuals’ levels of depth of reasoning. The mechanism designer thinks that all combinations of levels between 1 and $K$ are in principle possible. We do not specify the designer’s beliefs about how the individuals’ levels are distributed in $\{1, \ldots, K\}^I$. This is because our results are independent of the specific distribution of levels, as long as the upper bound $K$ is larger or equal to 2 and the distribution assigns positive probability on all profiles of levels in $\{1, \ldots, K\}^I$.\footnote{In fact, even weaker assumptions on the distribution of levels suffice for our results. For instance, the general necessary condition in Theorem 1 holds as long as for each individual $i$, the distribution of levels supports a profile $(k_i, k_{-i}) \in \{1, \ldots, K\}^I$ such that $k_i \geq 2$ and $k_j = k_i - 1$ for all $j \neq i$.} Taking $K = 1$ would mean that all participants have a depth of reasoning at most equal to 1, which seems rather implausible. Importantly, not being able to rule out the presence of as little as two levels of reasoning guarantees our conclusions, which also remain true in the presence of individuals with higher levels of depth of reasoning.

The mechanism $\mu$ implements up to level-$K$ the social choice function $f$ given the behavioral anchors $\alpha$ if (i) $S_i^{k_i}(\mu|\alpha)$ is nonempty, for all $i$ and $1 \leq k_i \leq K$, and (ii) $f = \mu \circ \sigma$, for all strategy profiles $\sigma$ such that, for each $i$, $\sigma_i \in S_i^{k_i}(\mu|\alpha)$ with $1 \leq k_i \leq K$. Part (ii) is the main restriction, requiring that the desired outcome prevails at all type profiles and independently of the strategies individuals follow, as long as they are consistent with the theory of level-$k$ reasoning for some depth of reasoning no greater than $K$. Part (i) rules out cases where (ii) is met only because of the absence of strategy profiles consistent with level-$k$ reasoning: best responses might not exist, for instance, in discontinuous mechanisms or when the message space is open.

We do not require implementability for $k_i = 0$. First, we think of all individuals as being minimally rational in the sense of playing a best response to some belief. In addition, this exclusion causes little loss of generality: the necessary condition for implementability derived in the next section, and the sufficient condition under truthful anchors derived in Section 5 hold when in-
cluding \( k_i = 0 \) in the definition as well. Intuitively, the planner accepts level-0 agents as a way to capture individuals’ gut feelings towards the mechanism, and hence, does not see herself as trying to affect those. The interesting problem of how to suggest or modify behavioral anchors might be of importance in a new direction of mechanism design, but it is beyond our scope here.

4 A General Level-\( k \) Revelation Principle

To understand the limits of level-\( k \) implementation, we start by showing how a slight strengthening of Bayesian incentive compatibility is necessary as soon as the social choice function is level-\( k \) implementable for some arbitrary behavioral anchors in any mechanism. This has two related and surprising implications. First, level-\( k \) reasoning does not free us from incentive compatibility constraints, even if the mechanism designer had the ability to choose the anchors in each mechanism. Second, incentive compatibility is a general necessary condition that will hold when studying level-\( k \) implementation, regardless of the regularity restrictions one is willing to place on behavioral anchors. Of course, such restrictions may generate supplementary necessary conditions, or turn necessary conditions into also sufficient, as we will see in later sections.

Say that a social choice function \( f \) is implementable up to level-\( K \) for some anchors if there exists a mechanism \( \mu \) and some behavioral anchors \( \alpha \) for \( \mu \) such that \( \mu \) implements up to level-\( K \) the social choice function \( f \) given \( \alpha \).

The next result may, at first glance, come as a surprise, as it shows that only the standard Bayesian incentive compatible social choice functions are implementable in this sense.

In fact, a slightly stronger property is necessary, with the incentive constraints being strict in some cases. There might be circumstances under which the mechanism designer wishes to implement a social choice function that is insensitive to some changes of an individual’s type. For instance, two types might differ only in higher-order beliefs, which may not matter to the mechanism designer for the problem at hand. For level-\( k \) implementation, incentive constraints need to be strict whenever comparing types for which the social
choice function is responsive. Formally, say that $f$ is *insensitive* when changing $i$’s type from $t_i$ to $t_i'$, denoted by $t_i \sim f^i t_i'$, if $f(t_i, t_{-i}) = f(t_i', t_{-i})$ for all $t_{-i}$. Otherwise, we say that $f$ is *responsive* to $t_i$ versus $t_i'$.

**Definition 1.** The social choice function $f$ is strictly-if-responsive Bayesian incentive compatible (SIRBIC) whenever (i) it is Bayesian incentive compatible, that is,

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

for all $t_i, t_i'$, and (ii) the inequality holds strictly when the social choice function is responsive to $t_i$ versus $t_i'$.

Our main result follows:

**Theorem 1.** Suppose $K \geq 2$. If a social choice function is implementable up to level-$K$ for some arbitrary anchors, then it satisfies SIRBIC.

**Proof.** Let $\mu$ be a mechanism that implements up to level-$K$ the social choice function $f$ given some behavioral anchors $\alpha = (\alpha_i)_{i \in I}$. For each $i$, let $\sigma_i^2$ be an element of $S_i^2(\mu|\alpha)$ (which is nonempty by definition of implementation up to level $K$ since $K \geq 2$).

We start by showing that $f$ is Bayesian incentive compatible. Consider two types $t_i$ and $t_i'$ in $T_i$. As $\sigma_i^2 \in S_i^2(\mu|\alpha)$, it follows that $\sigma_i^2$ is a best response for $i$ against some $\sigma_{-i}^1 \in S_{-i}^1(\mu|\alpha)$. We then have:

$$\int_{t_{-i} \in T_{-i}} u_i(f(t), t)dp(t_{-i}|t_i) = \int_{t_{-i} \in T_{-i}} u_i(\mu(\sigma_i^2(t_i), \sigma_{-i}^1(t_{-i})), t)dp(t_{-i}|t_i) \geq \int_{t_{-i} \in T_{-i}} u_i(\mu(\sigma_i^2(t_i'), \sigma_{-i}^1(t_{-i})), t)dp(t_{-i}|t_i) = \int_{t_{-i} \in T_{-i}} u_i(f(t_i', t_{-i}), t)dp(t_{-i}|t_i),$$

where the two equalities follow from the fact that $\mu$ implements $f$ up to level $K$ given the anchors $\alpha$, and the inequality follows from the fact that $\sigma_i^2(t_i)$ is one of $t_i$’s best responses against $\sigma_{-i}^1$. 

12
We establish the required strict inequalities with a reasoning by contraposition. Suppose that the incentive constraint for type $t_i$ pretending to be type $t'_i$ is binding. Then, the weak inequality in the previous paragraph must hold with equality, and the strategy $\tau_i$ belongs to $S^2_i(\mu|\alpha)$, where $\tau_i$ differs from $\sigma^2_i$ only in that $t_i$ picks $\sigma^2_i(t'_i)$.\footnote{$\tau_i$ is measurable as singletons in $T_i$ are measurable because $T_i$ is separable metrizable, and hence also Hausdorff.} By level-$k$ implementation, it must be that $f(t_i, t_{-i}) = \mu(\tau_i(t_i), \sigma^1_{-i}(t_{-i}))$ for all $t_{-i}$. This is equal to $\mu(\sigma^2_i(t'_i), \sigma^1_{-i}(t_{-i}))$, by definition of $\tau_i$, and to $f(t'_i, t_{-i})$, by definition of level-$k$ implementation. Hence, the social choice function must be insensitive when changing $i$’s type from $t_i$ to $t'_i$, which concludes the proof.

Theorem 1 can be viewed as a level-$k$ revelation principle, since SIRBIC is stronger than Bayesian incentive compatibility. For an intuition behind the result, note that each player’s level-$k$ strategy is a best response to other players’ level-$(k-1)$ strategies, but whenever $k > 1$, this level-$k$ strategy composed with those level-$(k-1)$ strategies must implement the social choice function, that is, due to the definition of implementation, the resulting outcome is the same as the truth-telling outcome under the social choice function. Thus, although players’ beliefs about other players’ strategies are not consistent, as they are in equilibrium, for $k > 1$, level-$k$ of each player consistently believes that it can at best get the truth-telling outcome under the social choice function. Since the truth-telling outcome is the best for him, then, in particular, lying when others tell the truth cannot be better. This is exactly Bayesian incentive compatibility (an additional step is required to take us to SIRBIC). It follows that one can decompose the strategy mappings for any such level-$k$ in the general mechanism into truth-telling in the direct mechanism and the transformation mapping from direct to indirect reporting.

### 4.1 Two Clarifying Examples

We next present two examples, which are illustrative of the power and limitations of our level-$k$ revelation principle. First, our Theorem 1 contrasts with
some more permissive results found in Crawford (2016). The following example from Crawford (2016) highlights the reasons underlying our contrasting results.

**Example 1.** Consider the bilateral trading problem with risk-neutral traders and one indivisible object. The buyer’s value $v$ and the seller’s cost $c$ are distributed uniformly on $[0, 1]$. They trade using the $\frac{1}{2}$-double auction. That is, the buyer and seller simultaneously submit respectively a bid and an ask. They trade the object if and only if the buyer’s bid is at least equal to the seller’s ask. If they do trade, then the trading price equals the average of the bid and ask. In case of trade, the buyer’s payoff is equal to her value minus the price while the seller’s payoff is equal to the price minus her cost. Each trader obtains zero whenever there is no trade.

Crawford (2016) assumes that the level-0 behavioral anchor is uniform random over $[0, 1]$. Then there is a unique level-1 consistent strategy, which equals bidding $\frac{2}{3}v$ for the level-1 buyer and asking $\frac{2}{3}c + \frac{1}{3}$ for the level-1 seller. This pair of strategies generates an outcome (which can be viewed as a social choice function) that is not incentive compatible. For instance, the buyer of value 0.5 expects a zero payoff under this outcome, and would be better off if she imitates the buyer of value 0.75. Thus, if all individuals are exclusively level-1 (more generally, homogeneous in their levels), then we can implement social choice functions that are not incentive compatible. Theorem 1 shows that if the designer has any doubt about this assumption of homogeneity, in that she cannot rule out that individuals may be of either level-1 or 2 (or possibly others above), then she is bound by the classic Bayesian incentive compatibility constraints. Such agnosticism is the norm in a mechanism design approach.

Let us then consider the situation where the individuals have heterogeneous levels. For instance, suppose traders could be either level-1 or level-2. Given the level-1 consistent strategies identified above, the following strategy is level-
2 consistent for the buyer

\[
\begin{align*}
\frac{2}{3}v + \frac{1}{9}, & \quad \text{if } v \geq \frac{1}{3} \\
\frac{1}{3}v, & \quad \text{if } v < \frac{1}{3},
\end{align*}
\]

while the following strategy is level-2 consistent for the seller

\[
\begin{align*}
\frac{2}{3}c + \frac{2}{9}, & \quad \text{if } c \leq \frac{2}{3} \\
c, & \quad \text{if } c > \frac{2}{3}.
\end{align*}
\]

Again, when paired up against one another, these level-2 consistent strategies result in an outcome that is not incentive compatible. For example, a buyer of valuation \( v = 1 \) bids \( 7/9 \), and hence, trades only with sellers with cost parameter \( c \in [0, 7/9] \). Her expected payoff would be strictly improved by imitating a buyer with valuation \( v = 5/6 \), bidding \( 2/3 \) instead, making her trade only with sellers whose cost parameter \( c \in [0, 2/3] \), but at a lower price.

The lack of incentive compatibility implies that the social choice function that results by pairing the two level-2 consistent strategies is nonimplementable in our sense. In fact, the buyer’s (seller’s) level-2 consistent strategy when paired with the seller’s (buyer’s) level-1 consistent strategy also results in an outcome that is not incentive compatible either. But notice that the four outcomes or social choice functions that are generated by the four pairs of buyer’s and seller’s levels are not equal to each other. Thus, when there are heterogeneous levels of players, then one way, and following Theorem 1, the only way, to get around incentive compatibility constraints is to implement different social choice functions for different levels of reasoning.\(^8\) However, such a differential treatment might be questionable on normative grounds because levels of reasoning do not determine individual preferences. 

The second example has implications that may take us far afield from the current paper. It shows that Bayesian incentive compatibility ceases to be

\(^8\)To be precise, this is true when the designer wants to implement social choice functions. As shown in Example 2, it is possible to get around incentive compatibility if the designer implements a social choice set.
necessary if the designer wants to implement a social choice set.⁹

**Example 2.** Suppose there are two individuals, and we wish to implement a social choice set \( \{f, f'\} \) such that the social choice function \( f \) is strictly Bayesian incentive compatible (strictly BIC) but \( f' \) is not BIC. Moreover, suppose that for both individuals, \( f' \) is uniformly worse than \( f \). That is, for all \( i \) and \( t \),

\[
    u_i(f(t), t) > \max_{\ell \in \{f'(t') : t' \in T\}} u_i(\ell, t).
\]

Consider the following mechanism \( \mu \): Each individual announces her type and one social choice function in \( \{f, f'\} \). Let \( t \) be the types reported by the individuals. If at least one individual announces \( f \), then the outcome is \( f(t) \) whereas if both individuals announce \( f' \), then the outcome is \( f'(t) \).

Suppose the behavioral anchor \( \alpha \) is such that level-0 announces her true type and \( f \). Then, given that \( f \) is strictly BIC, announcing one’s true type and either \( f \) or \( f' \) is a level-1 consistent strategy. Since \( f' \) is uniformly worse than \( f \), if a level-2 individual believes that the level-1 of the other individual will report her true type and announce \( f' \), the best response for the level-2 individual is to report her true type and announce \( f \). If a level-2 individual believes that the level-1 of the other individual will report her true type and announce \( f \), the best response for the level-2 individual is to report her true type and announce either \( f \) or \( f' \). Thus, announcing one’s true type and either \( f \) or \( f' \) is a level-2 consistent strategy. Iterating this argument, we obtain that announcing one’s true type and either \( f \) or \( f' \) is a level-\( k \) consistent strategy for all \( k \geq 1 \).

Then, it follows that irrespective of individuals’ levels \( k_i, k_j \geq 1 \),

\[
    \{\mu \circ \sigma : \sigma_i \in S_i^{k_i}(\mu|\alpha), \sigma_j \in S_j^{k_j}(\mu|\alpha)\} = \{f, f'\}.
\]

Thus, the above mechanism implements the social choice set at all combinations of levels, and yet \( f' \) is not BIC.

⁹Note that a social choice set instead of a social choice correspondence is the appropriate notion of set-valued rules in case of incomplete information.
5 Direct Mechanisms and Truthful Anchors

After having obtained a general level-$k$ revelation principle for arbitrary mechanisms and arbitrary behavioral anchors, the rest of the paper proceeds by investigating specific anchors and classes of mechanisms. Since level-$k$ reasoning has significantly different predictions than Nash equilibrium in many games, one might have thought that level-$k$ implementation would allow implementing social choice functions that are not weakly Nash implementable. We already saw in the previous section that this intuition is not correct. One may wonder now if level-$k$ implementation is not in fact much more restrictive than weak Nash implementation. This may depend on the stand one takes regarding behavioral anchors in the implementing mechanisms, but the rest of our results shows that there are important scenarios where SIRBIC is also sufficient for level-$k$ implementation.

In particular, this section uses truthful anchors in direct mechanisms. Experimental evidence offers support to their use. This is consistent with the well-known argument that truth-telling may be a focal or salient point. Also, even if the mechanism designer might not be able to nudge people to consider any anchor she would find convenient, making truth-telling salient enough to serve as the anchor may be easier. We now show that SIRBIC is sufficient for level-$k$ implementation via a direct mechanism with truthful anchors. We first state a lemma whose easy proof is left to the reader.

Lemma 1. Let $f$ be a social choice function. For each $i$, the relation $\sim_i^f$ is transitive: $t_i \sim_i^f t'_i$ and $t'_i \sim_i^f t''_i$, then $t_i \sim_i^f t''_i$. In addition, $f(t) = f(t')$ for any type profiles $t$ and $t'$ such that $t_i \sim_i^f t'_i$ for all $i \in I$.

Theorem 2. If $f$ satisfies SIRBIC, then for all $K \geq 1$, $f$ is implementable up to level-$K$ by a direct mechanism with truthful anchors.

Proof. The result can be proved by using $f$ itself as a direct mechanism. Let $\alpha^*$ denote the profile of truthful anchors. We begin with level-1 individuals.

---

10See, for example, Crawford (2003), Crawford and Iriberri (2007), Cai and Wang (2006), and Wang et al. (2010).
By Bayesian incentive compatibility, reporting \( t_i \) is a best response for \( i \) of type \( t_i \) against the truthful anchors for the other individuals. Reporting other types may be best responses as well, but only if the corresponding incentive constraint is binding. By SIRBIC, \( \sigma^1_i \) is a best response for \( i \) against the truthful anchors for the other individuals if and only if \( \sigma^1_i(t_i) \sim^D t_i \), for all \( t_i \). This characterizes \( S^1_i(f|\alpha^*) \). Since this holds for every \( i \), a simple application of Lemma 1 implies that \( f = f \circ \sigma \) for every \( \sigma \in S^1(f|\alpha^*) \).

Consider now a level-2 individual \( i \), who expects others to play \( \sigma^1_{-i} \in S^1_{-i}(f|\alpha^*) \). Her expected utility from reporting type \( t'_i \) when of type \( t_i \) is

\[
\int_{t_{-i} \in T_{-i}} u_i(f(t'_i, \sigma^1_{-i}(t_{-i})), t) dp(t_{-i}|t_i).
\]

By Lemma 1, this is equal to

\[
\int_{t_{-i} \in T_{-i}} u_i(f(t'_i, t_{-i}), t) dp(t_{-i}|t_i),
\]

which is the same as what \( t_i \) would get by such misreporting if others were truthful. Thus \( S^2_i(f|\alpha^*) = S^1_i(f|\alpha^*) \). In fact, using induction and the same argument, for all \( k \geq 2 \), \( S^k_i(f|\alpha^*) = S^1_i(f|\alpha^*) \). Lemma 1 then implies that, for all \( K \geq 1 \), \( f \) up to level-\( K \) implements \( f \) with truthful anchors. \( \square \)

We briefly observe that, if anchors are not truthful in a direct mechanism, then SIRBIC is sufficient for implementation up to level-\( K \) in the special case of independent private values when the type distribution coincides with the distribution of messages generated by the level-0 anchor (for example, when types are distributed uniformly and level-0 anchor in the direct mechanism is uniform random). Beyond this special case, SIRBIC and strict level-1 incentive compatibility (i.e., that truth-telling is the unique best reply to level-0) suffice for implementation up to level-\( K \). Level-1 incentive compatibility also features in Crawford (2016), for instance.
6 Augmented Mechanisms and Uniform/Atomless Anchors

In this section, we move beyond direct mechanisms. Uniform anchors, in the sense of picking an action uniformly at random, are often invoked in the literature, either to fit the behavior of experimental subjects in certain games, or more recently, in the context of implementation (Crawford (2016), or Gorelikina (2015) for uniform distribution environments). It is thus important to better understand level-$k$ implementation under uniform anchors. We provide sharp answers for this scenario as well. Perhaps even more surprising than for the case of truthful anchors, SIRBIC is also sufficient under uniform anchors, in an augmented mechanism that we construct, for continuous social choice functions in the case of independent private values. For more general belief environments, an additional necessary condition is identified and shown to be essentially sufficient, along with SIRBIC.

In our quest to draw conclusions that hold for a wide class of behavioral level-0’s, beyond the case of uniform anchors, the results in this section also hold under arbitrary atomless anchors whenever there is a continuum of messages, even if these anchors vary with types.\footnote{The sufficiency results in this section can be further extended to cover the case of mixed anchors that have an atom at truth-telling, while with the rest of probability, level-0 plays according to some arbitrary atomless distribution.}

6.1 Independent Private Values

Given a mechanism $\mu : M_1 \times \cdots \times M_I \to \Delta X$, the anchors $\alpha$ are uniform whenever, for each individual $i$, anchor $\alpha_i$ is the uniform probability distribution over $M_i$. Such anchors thus do not vary with types. More generally, assuming that $M_i$ contains a continuum of messages for each $i$, the anchors $\alpha$ are atomless if the distribution $\alpha_i(t_i)$ of messages contains no atom, for each $t_i$ and each $i$. For such mechanisms, atomless anchors are much more general than uniform anchors since they accommodate non-uniform distributions and the anchor can vary with types. One could imagine, for instance, that in
auctions anchors are biased to some extent towards truth-telling.

The environment satisfies private values if for all $i$, individual $i$’s Bernoulli utility function depends only on $i$’s type: $u_i(x, t) = u_i(x, t_i)$, for each $t$ and each $i$. Types are distributed independently if the prior can be written as the product of its marginals: $p = \prod_i p_i$, where $p_i$ denotes the marginal probability distribution on $T_i$. We maintain the following assumption for the rest of the paper:

**Assumption 1.** For all individuals $i$, the marginal distribution $p_i$ has full support.

Fix a social choice function $f$. An individual $i$ is irrelevant for $f$ if $f$ is insensitive to any change of types for $i$, that is, $t_i \sim_i t_i'$ for each $t_i, t_i' \in T_i$. Individuals who are not irrelevant are called relevant. Of course, by definition, the designer can determine whether an individual is relevant or irrelevant.

Consider now the following mechanism $\mu^f$. Each relevant individual reports a type along with a real number between 0 and 1. Let $i$’s report be $m_i = (t_i, z_i) \in T_i \times [0, 1]$ for each $i$. Then, the designer implements $f(t')$ where

$$t_i' = \begin{cases} 
\text{arbitrary } \bar{t}_i & \text{if } i \text{ is irrelevant} \\
t_i & \text{if } i \text{ is relevant and } z_i = 0 \\
drawn \text{ according to } p_i & \text{if } i \text{ is relevant and } z_i > 0.
\end{cases}$$

Here is our sufficiency result for environments with independent private values:

**Theorem 3.** Consider an environment with independent private values, and a social choice function $f$ that is continuous. If $f$ satisfies SIRBIC, then for all $K \geq 1$, $\mu^f$ implements $f$ up to level-$K$ given uniform anchors (or, more generally, atomless anchors).

**Proof.** The outcome being implemented does not depend on the types of irrelevant individuals. The mechanism designer thus need not consult them and can use without loss of generality any arbitrary type, for instance $\bar{t}_i$ for all
irrelevant $i$. For notational simplicity, we will assume from now on that all individuals are relevant.

Let $\alpha^U$ denote the uniform anchors (or, more generally, anchors that are atomless). We argue first that, for each individual $i$, $S_i^1(\mu^f|\alpha^U)$ is the set of reports $(\tau_i, 0)$ such that $\tau_i(t_i) \sim^f t_i$ for all $t_i$. Given the uniform anchors, such an individual $i$ of level 1 assigns zero probability to the event that others send a zero along with their type report. If individual $i$ picks a positive number along with some type report, then she expects the lottery

$$\int_{t \in T} f(t)dp(t).$$

(2)

If, on the other hand, she sends a zero along with some type report $t_i$, she expects the lottery

$$\int_{t_{-i} \in T_{-i}} f(t_i, t_{-i})dp_{-i}(t_{-i}).$$

(3)

Suppose now that individual $i$’s type is $t^*_i$. Her expected utility under lottery (3) is

$$u_i\left(\int_{t_{-i} \in T_{-i}} f(t_i, t_{-i})dp_{-i}(t_{-i}), t^*_i\right) = \int_{t_{-i} \in T_{-i}} u_i(f(t_i, t_{-i}), t^*_i)dp_{-i}(t_{-i}).$$

By SIRBIC, we have

$$\int_{t_{-i} \in T_{-i}} u_i(f(t^*_i, t_{-i}), t^*_i)dp_{-i}(t_{-i}) \geq \int_{t_{-i} \in T_{-i}} u_i(f(t_i, t_{-i}), t^*_i)dp_{-i}(t_{-i}),$$

(4)

for all $t_i$, with a strict inequality for all $t_i$ such that $t_i \not\sim^f t^*_i$.

Since $f$ is continuous, $i$ is relevant, and $p_i$ has full support, there is a positive $p_i$-measure of $t_i$’s types for which inequality (4) holds strictly. Using this observation, we keep a strict inequality when integrating (4) over $t_i$:

$$\int_{t_{-i} \in T_{-i}} u_i(f(t^*_i, t_{-i}), t^*_i)dp_{-i}(t_{-i}) > \int_{t \in T} u_i(f(t), t^*_i)dp(t),$$

which is equal to the expected utility of lottery (2). Thus, sending a type along
with a positive number is never a best response against the uniform anchors, since sending \((t^*_i, 0)\) is strictly better.

Among reports that include a zero, truthfully reporting one's type is a best response, by (4), and so is any type \(t_i \sim t^*_i\). Reporting types \(t_i \not\sim t^*_i\), however, is strictly inferior. Thus we have proved, as claimed, that \(S^1_i(\mu^f|\alpha^U)\) is the set of reports \((\tau_i, 0)\), where \(\tau_i(t_i) \sim t_i\) for all \(t_i\).

We now show that \(S^k_i(\mu^f|\alpha^U) = S^1_i(\mu^f|\alpha^U)\), for all \(i\) and all \(k \geq 2\). This will conclude the proof that for all \(K \geq 1\), \(\mu^f\) up to level-\(K\) implements \(f\) with uniform anchors, thanks to Lemma 1. Level-2 of individual \(i\) believes that level-1 of any individual \(j\) plays according to strategies in \(S^1_i(\mu^f|\alpha^U)\). As already argued in the proof of Theorem 2, Lemma 1 implies that we can assume without loss of generality that individual \(j\)'s type report is truthful (because nontruthful reports result in the same outcome by definition of \(\sim^f\)). Thus, individual \(i\) expects the lottery (2) if she sends a positive number along with her type report, and lottery (3) if she sends zero along with a type report \(t_i\). These are the same lotteries as for our level-1 reasoning, but for a different reason, namely because others are now expected to send a truthful type report with a zero. The comparison of these two lotteries remains unchanged, and we get \(S^2_i(\mu^f|\alpha^U) = S^1_i(\mu^f|\alpha^U)\). The argument extends trivially to any higher depth of reasoning \(k > 2\). \(\square\)

Sufficiency of SIRBIC is determined only for the case of continuous social choice functions. We see continuity as a mild requirement that is always satisfied, for instance, in the case of finite type sets. In the presence of a continuum of types, many SIRBIC social choice functions can be approximated by continuous social choice functions that satisfy SIRBIC as well. We have identified weaker conditions under which SIRBIC remains sufficient,\(^{12}\) but finding a necessary and sufficient condition for level-\(k\) implementation with uniform anchors remains an open question on the class of all social choice functions.

\(^{12}\)Indeed, there are ways to dispense with continuity as well as Assumption 1, by assuming that all individuals are relevant in a slightly stronger sense: for all \(i\), there exists a \(t_{-i}\) such that all \(f\)-equivalent classes at \(t_{-i}\) have less than probability one. An \(f\)-equivalent class at \(t_{-i}\) is an element of the partition of \(T_i\) generated by the equivalence relation \(\sim\) on \(T_i\), where \(t_i \sim t'_i \iff f(t_i, t_{-i}) = f(t'_i, t_{-i})\).
The social choice function $f$ is used in $\mu^f$ as if in a direct mechanism when the designer takes type reports into account. SIRBIC essentially guarantees that truth-telling is the only best response to truth-telling (up to the equivalence relations $\sim_i^f$). Using $f$ as a direct mechanism (as for Theorem 2) would not work, though, because level-1 individuals would usually not have the right expectations (unless they had a uniform prior). The mechanism $\mu^f$ succeeds by effectively separating individuals’ beliefs when having a depth of reasoning 1 or 2+. A level-1 individual expects that others will submit a positive number, in which case the mechanism proceeds so as to have this individual face the same expected outcome under $f$ as if others where truth-telling. A level 2+ individual expects that others will submit a zero, in which case the type report is taken into account and $f$ is used to compute the outcome. Crucially, individuals never wish to report a positive number, whatever their depth of reasoning $k \geq 1$, because the SIRBIC inequalities are preserved under averages.

In many classic implementation problems, including simple auctions and bilateral trade problems (also studied by Crawford (2016)), type sets are intervals. In such cases, any SIRBIC social choice function can be level-$k$ implemented given uniform anchors by a direct mechanism. This follows at once from the last result after observing that there always exists an isomorphism between $T_i$ and $T_i \times [0,1]$ in such cases. However, it is possible to construct examples where simply using the social choice function itself as a direct mechanism does not work, and examples with finite types sets where one must use an indirect mechanism to implement the social choice function.

In this section, we take the view that anchors are uniform/atomless independently of the mechanism in use. This makes sense if individuals’ gut reaction to a game is totally random. This would be the case, for instance, if they fail to completely grasp an understanding of the link between actions and outcomes. We find it plausible, though, that different games may trigger different anchors. Reporting zero when participating in $\mu^f$ may be salient enough that anchors would display an atom at zero. However, in the spirit of framing effects, it is also possible that other, perhaps less transparent, de-
criptions of \( \mu^f \) would make atomless anchors more likely.\(^{13}\) Interestingly, we note that a modified mechanism in which the role played by the number zero in \( \mu^f \) is given to a finite (or countably infinite) set Zero of numbers (Zero= \( \{0, \ldots, n'/n, \ldots, 1\} \), with integers \( n' < n \), where one can choose how fine the grid \( n \) is arbitrarily) would give the same result. Perhaps with such a modification, uniform or atomless anchors may seem more plausible to more individuals. Whether individuals’ behavior is best described using uniform anchors when participating in \( \mu^f \), or other related mechanisms, is an interesting empirical question that goes beyond the scope of this paper.

Further study of how games and their description might impact anchors is a fascinating topic that is not yet well-understood. Progress on that front will then have to be incorporated into the theory of level-\( k \) implementation. Notice, though, that Theorem 1 holds in this more general model as well, and that SIRBIC thus remains necessary.

6.2 The General Case beyond IPV

In the absence of independent private values, SIRBIC need not be sufficient anymore for level-\( k \) implementation given uniform anchors. The next example and result show this.

**Example 3.** Suppose that \( X = \{x, y\} \), \( T_1 = T_2 = \{a, b\} \), \( p \) is uniform, and there is pure common interest, with the following dichotomous Bernoulli utility functions:

\[
\begin{align*}
    u_i(x, t) &= 1 \text{ and } u_i(y, t) = 0 \text{ for } t = (a, a) \text{ or } (b, b) \\
    u_i(y, t) &= 1 \text{ and } u_i(x, t) = 0 \text{ for } t = (a, b) \text{ or } (b, a)
\end{align*}
\]

The Pareto social choice function that picks \( x \) if \( (a, a) \) or \( (b, b) \), and \( y \) otherwise, satisfies SIRBIC. Using it as a direct mechanism does not allow to level-\( k \)

---

\(^{13}\) That different descriptions of the same mechanism may impact realized outcomes and implementability is absent when individuals are rational. Glazer and Rubinstein (2014) is the first paper investigating this new feature for a different notion of bounded rationality.
implement it given uniform anchors, as a level-1 individual expects the same lottery ($x$ or $y$ with equal probability) when reporting $a$ or $b$. One might conjecture that the Pareto social choice function could be implemented via an indirect mechanism. This is not the case, though, as we will show after the next theorem.

The next theorem identifies an additional necessary condition for level-$k$ implementation, while the theorem that follows will identify a large class of problems where it becomes sufficient once combined with SIRBIC when there are at least three (or just one) relevant individuals. The case of exactly two relevant individuals is discussed at the end. The necessary result holds very generally (as long as anchors are type-independent), while the sufficiency result extends to settings when anchors are atomless in mechanisms with a continuum of messages.

Individual $i$’s (interim) preference over state-independent or constant lotteries, i.e., over $\Delta(X)$, for type $t_i$ is given by:

$$U_i(\ell|t_i) = \int_{t_-i}^{t_+i} u_i(\ell, t) dp(t_-i|t_i).$$

Individual $i$ has different preferences over constant lotteries at $t_i$ and $t'_i$ if there does not exist $\alpha > 0$ and $\beta$ such that $U_i(\cdot|t_i) = \alpha U_i(\cdot|t'_i) + \beta$. The following condition is a stronger version of a condition that first appeared under the name of measurability in Abreu and Matsushima’s (1992) paper on virtual implementation in iteratively undominated strategies under incomplete information. A-M measurability is defined with respect to a partition of the type space that results after an iterative process of type separation, as a function of their interim preferences over increasingly enlarged classes of lotteries. Our condition corresponds to the first step of that iterative process.

**Definition 2.** The social choice function $f$ is first-step A-M measurable (FSAMM) whenever $t_i \not\sim f t'_i$ implies that individual $i$ has different interim preferences over constant lotteries at $t_i, t'_i$.

An additional necessity result is provided for these more general settings:
**Theorem 4.** If a social choice function $f$ is implementable up to level $K$ given uniform anchors (or, more generally, type-independent anchors), then $f$ is FSAMM.

**Proof.** Let $\mu$ be a mechanism that implements $f$ up to level $K$ given uniform anchors $\alpha^U$. For each individual $i$, let $\sigma_i^1$ be some level-1 consistent strategy, that is, $\sigma_i^1 \in S_i^1(\mu|\alpha^U)$. For each type $t_i$, let $\ell_i(t_i)$ be the lottery over $X$ that a level-1 individual $i$ expects to occur when playing $\sigma_i^1$. Formally,

$$\ell_i(t_i) = \int_{m_{-i} \in M_{-i}} \mu(\sigma_i^1(t_i), m_{-i}) d\alpha^U_{-i}(m_{-i}).$$

Suppose that individual $i$'s interim preference over constant lotteries is the same when of type $t_i$ as when of $t_i'$. Lottery $\ell_i(t_i')$ is the best lottery she can get by reporting a message in the mechanism when of type $t_i'$. Hence it is also the best lottery she can get by reporting a message in the mechanism when of type $t_i$. The strategy $\tau_i$ that coincides with $\sigma_i^1$ except that $\tau_i(t_i) = \tau_i(t_i') = \sigma_i^1(t_i')$ then also belongs to $S_i^1(\mu|\alpha^U)$. By definition of implementability, $f(t_i, t_{-i}) = \mu(\tau_i(t_i), \sigma_{-i}^1(t_{-i}))$ and $f(t_i', t_{-i}) = \mu(\tau_i(t_i'), \sigma_{-i}^1(t_{-i}))$ for all $t_{-i}$. But since $\tau_i$ picks the same message for $t_i$ and $t_i'$, we have $t_i \sim f t_i'$. Hence, $f$ is FSAMM.

Returning to Example 3, note how both types of each agent have identical interim preferences over constant lotteries. Thus, FSAMM would require that the social choice function be constant over all states, and clearly, the Pareto function is not. Therefore, this function is not level-$K$ implementable given uniform or type-independent anchors.

Under independent private values, FSAMM is implied by SIRBIC. To see this, suppose that individual $i$'s interim preference over constant lotteries is the same when of type $t_i$ as when of $t_i'$. By contradiction, suppose that $t_i \not\sim f t_i'$. By SIRBIC,

$$\int_{t_{-i} \in T_{-i}} u_i(f(t_i, t_{-i}), t_i) dp_{-i}(t_{-i}) > \int_{t_{-i} \in T_{-i}} u_i(f(t_i', t_{-i}), t_i) dp_{-i}(t_{-i}),$$

26
and
\[ \int_{t \in T} u_i(f(t', t_{-i}), t') dp_{-i}(t_{-i}) > \int_{t \in T} u_i(f(t, t_{-i}), t') dp_{-i}(t_{-i}). \]

Define lotteries \( \ell_i(t_i) = \int_{t \in T} f(t, t_{-i}) dp_{-i}(t_{-i}) \) and \( \ell_i(t'_i) = \int_{t \in T} f(t', t_{-i}) dp_{-i}(t_{-i}) \). The two inequalities imply that \( u_i(\ell_i(t_i), t_i) > u_i(\ell_i(t'_i), t_i) \) and \( u_i(\ell_i(t'_i), t'_i) > u_i(\ell_i(t_i), t'_i) \), contradicting that these two types have the same preferences over constant lotteries.

Consider a social choice function \( f \) for an environment where type sets are finite. For each relevant individual \( i \), define the function \( \ell_i : T_i \rightarrow \Delta X \) such that individual \( i \) of type \( t_i \) weakly prefers \( \ell_i(t_i) \) over \( \ell_i(t'_i) \) for each \( t'_i \), and strictly prefers \( \ell_i(t_i) \) over \( \ell_i(t''_i) \), for each \( t''_i \) such that \( t''_i \not\sim f t_i \). Such a function always exists under FSAMM if the environment satisfies a weak condition of no-total-indifference, i.e., for all types \( t_i \) and individuals \( i \), the interim preferences \( U_i(\cdot|t_i) \) are such that \( t_i \) is never completely indifferent over all alternatives in \( X \) (the reader is referred to Abreu and Matsushima (1992, Lemma 1) or Serrano and Vohra (2005, Lemma 1) for the technical details of similar results).

Consider now the following mechanism \( \nu^f \). As in \( \mu^f \), each relevant individual reports a type along with a real number between 0 and 1. Letting \( I^* \) denote the set of relevant individuals, and assuming that there are \( r \geq 3 \) of them, the outcome under \( \nu^f \) is then determined as follows:

- If all relevant individuals submit a strictly positive number along with their type report, then the mechanism designer randomizes uniformly among relevant individuals, and picks the personalized lottery \( \ell_j(t_j) \) for the selected individual \( j \) for the type \( t_j \) picked at random following \( p_j \). In other words, the outcome is the lottery
  \[ \frac{1}{r} \sum_{j \in I^*} \sum_{t \in T_j} \ell_j(t_j)p_j(t_j). \] (5)

- If all but one relevant individuals - say \( i \) - submit a strictly positive
number along with their type report, then the mechanism designer picks the same lottery as above, with the only exception that \( i \)'s personalized lottery is the one associated to his type report instead of being chosen at random. In other words, the outcome is the lottery

\[
\frac{1}{r} \left( \ell_i(t_i) + \sum_{j \in I \setminus \{i\}} \sum_{t_j \in T_j} \ell_j(t_j)p_j(t_j) \right),
\]

(6)

where \( t_i \) is \( i \)'s type report.

- In all other cases, \( \nu^f \) coincides with \( \mu^f \).

This is our next sufficiency result:

**Theorem 5.** Suppose that type sets are finite, the environment satisfies no-total-indifference, and that there are at least three relevant individuals. If the social choice function \( f \) satisfies SIRBIC and FSAMM, then for all \( K \geq 1 \), \( \nu^f \) implements \( f \) up to level-\( K \) given uniform anchors (or, more generally, atomless anchors).

**Proof.** Again without loss of generality and for notational simplicity, we assume in the proof that all individuals are relevant. Let \( \alpha^U \) denote the uniform anchors (or, more generally, anchors that are atomless). We argue first that, for each individual \( i \), \( S^1_i(\nu^f|\alpha^U) \) is the set of reports \((\tau_i,0)\) such that \( \tau_i(t_i) \sim^f t_i \) for all \( t_i \). Given the uniform anchors, such an individual \( i \) of level 1 assigns zero probability to the event that others send a zero along with their type report. Recall that FSAMM and no-total-indifference yields the existence of the menu of lotteries \( \ell_i : T_i \to \Delta X \). If individual \( i \) picks a positive number along with some type report, then she expects the lottery (5). If, on the other hand, she sends a zero along with some type report \( t_i \), she expects the lottery (6). Suppose now that individual \( i \)'s type is \( t^*_i \). By linearity of \( i \)'s interim preference \( U_i(\cdot|t^*_i) \) when of type \( t^*_i \), her expected utility under lottery (6) is equal to

\[
\frac{1}{r} \left( U_i(\ell_i(t_i)|t^*_i) + \sum_{j \in I \setminus \{i\}} U_i(\sum_{t_j \in T_j} \ell_j(t_j)p_j(t_j)|t^*_i) \right),
\]
while her expected utility under lottery (5) is equal to
\[
\frac{1}{r} \left( U_i\left(\sum_{t_i \in T_i} \ell_i(t_i)p_i(t_i)|t_i^*\right) + \sum_{j \in T \setminus \{i\}} U_i\left(\sum_{t_j \in T_j} \ell_j(t_j)p_j(t_j)|t_i^*\right)\right).
\]

One of the best lotteries for type \( t_i^* \) that can be obtained when reporting a zero – getting a lottery as in (6) – is thus obtained by picking \( t_i = t_i^* \) since \( U_i(\ell_i(t_i^*)|t_i^*) \geq U_i(\ell_i(t_i)|t_i^*) \) for all \( t_i \), by definition of \( \ell_i \). Remember also that this inequality is strict for all \( t_i \) such that \( t_i \not\sim t_i^* \). The same argument as in the proof of Theorem 3 can be used to assert that the inequality still holds strictly when integrating with respect to \( t_i \) on both sides:
\[
U_i(\ell_i(t_i^*)|t_i^*) > U_i\left(\sum_{t_i \in T_i} \ell_i(t_i)p_i(t_i)|t_i^*\right).
\]

Hence, reporting a strictly positive number is not a best response for \( i \) of type \( t_i^* \) against uniform anchors, since reporting \( (t_i^*, 0) \) gives a strictly higher expected payoff, and a report \( (t_i, 0) \) is a best response if and only if \( t_i \sim t_i^* \).

The rest of the proof is the same as the proof of Theorem 3 because of SIRBIC and the fact (which follows from the step just proved) that \( \nu \) coincides with \( \mu \) in the case relevant for computing \( S_k(\nu|\alpha U) \) for \( k \geq 2 \).

The proof of this result and that of Theorem 3 offer some similarities as well as some differences. First, the lotteries \( \ell_i : T_i \to \Delta X \) that can be found thanks to FSAMM and no-total-indifference are used to ensure that reporting the true type along with the number zero is the only best reply to atomless beliefs (up to \( f \)-equivalent types). Once this is established, the social choice function \( f \) is used, as in \( \mu \), as if in a direct mechanism when the designer takes type reports into account. In that part of the argument, SIRBIC again guarantees that truth-telling is the only best response to truth-telling (up to the equivalence relations \( \sim \)).

The same mechanism also works for the case of only one relevant individual. If this were the case, at the beginning of the proof, she would have to make the comparison of lotteries (5) and (6), arriving at the same conclusion. The case
of exactly two relevant individuals is a bit more tricky. The difficulty arises when exactly one of the relevant individuals reports the number zero and the other a positive number. The mechanism \( \nu^f \) is not well defined in this case, as it would use \( f \) as well as the \( \ell_i \) to determine the outcome. While we have not worked out the details, we conjecture that a more involved mechanism that would randomize between \( f \) and the \( \ell_i \)’s, much along the lines of the literature on virtual implementation, should do the job for this case.

7 Concluding Remarks

1. We presented our results under the assumption that individuals see others’ levels of depth of reasoning as exactly one level below theirs. While this is one of the standard specifications, one can certainly envision more general scenarios. All our results can easily be adapted to a wide class of theories where individuals see others as less sophisticated as themselves. This would include, for instance, all the theories described through the language of cognitive hierarchies (Strzalecki’s (2014)), which subsumes earlier models by Stahl (1993), Stahl and Wilson (1994, 1995), and Camerer et al. (2004) among others.

2. Implementation in our sense is quite flexible, as the model can accommodate a wide variety of reasonings (and thus behaviors) as discussed earlier. While related to rationalizable full implementation, also with an iterative construction, our definition is less demanding, as individuals’ depth of reasoning is bounded and behavior at cognitive state of depth 0 is fixed. Bergemann et al. (2011) studies rationalizable implementation of social choice functions, and Kunimoto and Serrano (2016) consider correspondences. The diverging conclusions of these two papers, in terms of the permissiveness of the results, should bring a word of caution. Having restricted attention in this paper to social choice functions as a natural first step, we find it an interesting research agenda to investigate set-valued rules instead, and to figure out in particular whether implementation can be significantly different when behavior is better described via level-\( k \) than via Bayesian Nash equilibrium (recall Example 2).

3. Finally, assuming that behavior is governed by bounded levels of rea-
sioning leads in this paper to restoring a restrictive result. That is, even in such contexts, one cannot ignore the constraints imposed by Bayesian incentive compatibility. This is in marked contrast with the permissive implications that allowing such unsophisticated behavior has in the problem of continuous implementation, as shown in de Clippel et al. (2015). That is, if one insists on implementation being performed in a continuous manner, stronger versions of Maskin monotonicity, which can be very restrictive, have been found to be required on top of the incentive constraints if one insists on equilibrium logic (Oury and Tercieux (2012)). And yet, as shown in de Clippel et al. (2015), continuous implementation with bounded levels of reasoning relies only on the incentive constraints. It is therefore remarkable that incentive compatibility raises its stature, to describe the limits of decentralization, with or without continuity, once one abandons the notion of rational expectations.
References


