Level-k Mechanism Design

Geoffroy de Clippel
Brown University

Rene Saran
Yale-NUS College

Roberto Serrano
Brown University

This version: July 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 mechanism design, however, they turn out to have surprisingly similar implications. This paper provides tight necessary and sufficient conditions for implementation with bounded depth of reasoning, discussing the role and implications of different behavioral anchors. The central condition slightly strenghthens standard incentive constraints, and we term it strict-if-responsive Bayesian incentive compatibility (SIRBIC).

JEL Classification: C72, D70, D78, D82.

Keywords: mechanism design; bounded rationality; level k reasoning; revelation principle; incentive compatibility.

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 has shown that theories of level-k choice may provide a better description of people's behavior, especially when they are inexperienced.[1]The point this paper makes, however, is that when it comes

[1]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).

to mechanism design, the two approaches turn out to have surprisingly similar implications.

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. However, a host of bounded rationality notions have been put to the task.[2]

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.

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 social choice 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

[2]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, 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.

level-k reasoning model, in which each level-k individual best-responds to her belief that all her opponents are of level-$(k-1)$.[3]

In any theory that relies on bounded levels of reasoning, predictions depend on how one sets the anchor. Whether the mechanism designer can impact the anchor is debatable. Implementation is most permissive when giving her the freedom to pick the anchor. Thus, it comes perhaps 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 first 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 *strict-if-responsive Bayesian incentive compatibility* (SIRBIC). Theorem 1 thus asserts that if a social choice function is implementable up to level-K for $K \geq 2$ and for *any* behavioral anchor at level 0, it must satisfy SIRBIC. The exact converse holds, as shown in our next result (Theorem 2). The applicability of this converse 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.

It is appropriate to dwell on the importance of using a variety of behavioral anchors for level-0. In particular, the literature (e.g., Crawford's (2016) level-k analysis of the classic bilateral trading problem in Myerson and Satterthwaite (1983)) has discussed the use of uniform behavioral anchors, in which the gut reaction to a mechanism is to play an action chosen uniformly randomly from the available actions. We include a section with general results for uniform anchors. These findings confirm the theme of our general results. With independent private values, SIRBIC alone suffices for level-k implementation of continuous social choice functions under uniform anchors (Theorem 3).[4] Beyond independent private values, an additional weak necessary condition is uncovered, which amounts to the social choice reacting to types having differ-

[3]We observe in the concluding remarks section that our results apply to many other specifications of bounded levels of reasoning.

[4]Continuity can be dispensed with, as we discuss later.

ent interim preferences (Theorem 4). Conversely, this measurability condition and SIRBIC are also sufficient even if anchors are uniform (Theorem 5).

The take-away message of our work is that incentive compatibility arises as the robust condition that is key to describe the scope of implementable rules, even if bounded levels of reasoning are factored into the model. In principle, one might have thought that relaxing the requirement of Nash equilibrium would allow the planner to implement a wider set of goals, but, as we show, if one does not unduely restrict the levels of reasoning allowed in the model, that turns out to be a false hope.

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." Section 5 presents our sufficiency results for truthful anchors used in direct mechanisms. Section 6 contains our treatment of uniform 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. 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 ΔX 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 ΔM_i is the set of probability distributions over M_i. A strategy profile σ and type profile t induce a lottery $\mu(\sigma(t))$ over X.[5]

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 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 μ 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.

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* α_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 $\alpha = (\alpha_i)_{i \in I}$. 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 anchors, that is, $S_i^1(\mu|\alpha)$ is the set of strategies σ_i such that $\sigma_i(t_i)$ maximizes $\int_{t_{-i}} u_i(\mu(m_i, \alpha_{-i}(t_{-i})), t) dp(t_{-i}|t_i)$ over $m_i \in M_i$. By induction, for each $k \geq 1$,

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

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, $S_i^{k+1}(\mu|\alpha)$ is the set of strategies σ_i such that $\sigma_i(t_i)$ maximizes $\int_{t_{-i}} u_i(\mu(m_i, \sigma_{-i}(t_{-i})), t)dp(t_{-i}|t_i)$, for some $\sigma_{-i} \in S_{-i}^k(\mu|\alpha)$. 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 reasoning 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' depths of reasoning. The mechanism designer thinks that all combinations of levels in $\{1, \ldots, K\}$ are in principle possible. Our results are robust in the sense of being independent of K, as long as it is larger or equal to 2. 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 depths of reasoning.

The mechanism μ *implements up to level-*K the social choice function f given the behavioral anchors α 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 σ 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. Depths of reasoning are allowed to vary in the population. 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 including $k_i = 0$ 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 Necessary Condition

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 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, as we will see in Subsection 6.2.

Say that a social choice function f is *implementable up to level-K for some anchors* if there exists a mechanism μ and some behavioral anchors α for μ such that μ implements up to level-K the social choice function f given α. 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 permissive 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 mecha-

nism 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 $t_i \sim_i^f t_i'$, 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,

$$\int_{t_{-i}\in T_{-i}} u_i(\mu(t), t)dp(t_{-i}|t_i) \geq \int_{t_{-i}\in T_{-i}} u_i(\mu(t_i', t_{-i}), t)dp(t_{-i}|t_i), \qquad (1)$$

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 first result follows:

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

Proof. Let μ 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 σ_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 σ_i^2 is a best response for i against some $\sigma_{-i}^1 \in S_i^1(\mu|\alpha)$. We then have:

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

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

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 τ_i belongs to $S_i^2(\mu|\alpha)$, where τ_i differs from σ_i^2 only in that t_i picks $\sigma_i^2(t_i')$.[6] By level-k implementation, it must be that $f(t_i, t_{-i}) = \mu(\tau_i(t_i), \sigma_{-i}^1(t_{-i}))$ for all t_{-i}. This is equal to $\mu(\sigma_i^2(t_i'), \sigma_{-i}^1(t_{-i}))$, by definition of τ_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. □

We use SIRBIC because, with it, we are able to close the gap between necessary and sufficient conditions, which is of course important in order to best understand any notion of implementability. Since SIRBIC is stronger than Bayesian incentive compatibility, Theorem 1 can be viewed as a level-k revelation principle, and it contrasts with some more permissive results found in Crawford (2016). Note, though, how in order to generate level-k implementability beyond the constraints imposed by Bayesian incentive compatibility, that paper considers examples where only level-2 or only level-1 agents are present. Theorem 1 shows that if the planner has any doubt about this assumption, 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. Interestingly, the necessary condition, which is independent of K, is also sufficient: this will follow at once from Theorem 2 in the next section.

[6] τ_i is measurable as singletons in T_i are measurable because T_i is separable metrizable, and hence also Hausdorff.

5 Truthful Anchors

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, but the rest of the paper shows that there are important scenarios where SIRBIC is also sufficient for level-k implementation.

Experimental evidence offers support to the use of truthful anchors in direct mechanisms,[7] which 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 α^T denote the profile of truthful anchors. We begin with level-1 individuals. 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

[7] See, for example, Crawford (2003), Crawford and Iriberri (2007), Cai and Wang (2006), and Wang *et al.* (2010).

constraint is binding. By SIRBIC, σ_i^1 is a best response for i against the truthful anchors for the other individuals if and only if $\sigma_i^1(t_i) \sim_i^f t_i$, for all t_i. This characterizes $S_i^1(f|\alpha^T)$. Since this holds for every i, a simple application of Lemma 1 implies that $f = f \circ \sigma$ for every $\sigma \in S^1(f|\alpha^T)$.

Consider now a level-2 individual i, who expects others to play $\sigma_{-i}^1 \in S_{-i}^1(f|\alpha^T)$. Her expected utility from reporting type t_i' when of type t_i is

$$\int_{t_{-i} \in T_{-i}} u_i(f(t_i', \sigma_{-i}^1(t_{-i})), t) dp(t_{-i}|t_i).$$

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 $S_i^2(f|\alpha^T) = S_i^1(f|\alpha^T)$. In fact, using induction and the same argument, for all $k \geq 2$, $S_i^k(f|\alpha^T) = S_i^1(f|\alpha^T)$. Lemma 1 then implies that, for all $K \geq 1$, f up to level-K implements f with truthful anchors. □

We briefly observe that, if anchors are not truthful in a direct mechanism, 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).

6 Uniform and Atomless Anchors

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)). 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 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.

It is important to remark that, beyond the case of uniform anchors, results in this section also hold under arbitrary atomless anchors whenever there is a continuum of messages, even if these anchors vary with types.

6.1 Independent Private Values

Given a mechanism $\mu : M_1 \times \cdots \times M_I \to \Delta X$, the anchors α are uniform whenever, for each individual i, anchor α_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 α 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, $t_i \sim_i^f t_i'$ 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 μ^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

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

Here is our sufficiency result for independent private values environments:

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 $K \geq 1$, μ^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 α^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 $\tau_i(t_i) \sim_i^f t_i$ 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

$$\int_{t \in T} f(t) dp(t). \tag{2}$$

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

$$\int_{t_{-i} \in T_{-i}} f(t_i, t_{-i}) dp_{-i}(t_{-i}). \tag{3}$$

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

$$\int_{t_{-i}\in T_{-i}} u_i(f(t_i^*,t_{-i}),t_i^*)dp_{-i}(t_{-i})\geq\int_{t_{-i}\in T_{-i}} u_i(f(t_i,t_{-i}),t_i^*)dp_{-i}(t_{-i}),\qquad(4)$$

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

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 $t_i \sim_i^f t_i^*$. Reporting types $t_i \not\sim_i^f t_i^*$, however, is strictly inferior. Thus we have proved, as claimed, that $S_i^1(\mu^f|\alpha^U)$ is the set of reports $(\tau_i,0)$, where $\tau_i(t_i)\sim_i^f t_i$ for all t_i.

We now show that $S_i^k(\mu^f|\alpha^U)=S_i^1(\mu^f|\alpha^U)$, for all i and all $k\geq 2$. This will conclude the proof that for all $K\geq 1$, μ^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 $S_i^1(\mu^f|\alpha^U)$. 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 $\sim_i^f$). 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 $S_i^2(\mu^f|\alpha^U) = S_i^1(\mu^f|\alpha^U)$. The argument extends trivially to any higher depth of reasoning $k > 2$. □

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,[8] 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.

The social choice function f is used in μ^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 μ^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 in-

[8]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 $t_i \sim t_i' \iff f(t_i, t_{-i}) = f(t_i', t_{-i})$.

tervals. 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 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 μ^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, descriptions of μ^f would make atomless anchors more likely.[9] Interestingly, we note that a modified mechanism in which the role played by the number zero in μ^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 μ^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.

[9]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.

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 1. 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:

$$u_i(x, t) = 1 \text{ and } u_i(y, t) = 0 \text{ for } t = (a, a) \text{ or } (b, b)$$

$$u_i(y, t) = 1 \text{ and } u_i(x, t) = 0 \text{ for } t = (a, b) \text{ or } (b, a)$$

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 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 given uniform anchors, 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 more generally (when anchors are type-independent), while the sufficiency result extends to more 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 $\Delta(X)$, when of type t_i is given by:

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

Individual i has different preferences over constant lotteries at t_i and t_i' if there does not exist $\alpha > 0$ and β 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 $t_i \not\sim_i^f t_i'$ 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 μ be a mechanism that implements f up to level K given uniform anchors α^U. For each individual i, let σ_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 σ_i^1. Formally,

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

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 τ_i that coincides with σ_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 τ_i picks the same message for t_i and t_i', we have $t_i \sim_i^f t_i'$. Hence, f is FSAMM. □

Returning to Example 1, 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 $t_i \not\sim_i^f t_i'$. By SIRBIC,

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

and

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

Define lotteries $\ell_i(t_i) = \int_{t_{-i}} f(t_i,t_{-i})dp_{-i}(t_{-i})$ and $\ell_i(t_i') = \int_{t_{-i}} f(t_i',t_{-i})dp_{-i}(t_{-i})$. 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 $\ell_i : T_i \to \Delta X$ 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 $f(t_i,t_{-i}) \neq f(t_i'',t_{-i})$ for some 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 ν^f. As in μ^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 ν^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

$$\frac{1}{r}\sum_{j\in I^*}\sum_{t_j\in T_j}\ell_j(t_j)dp_j(t_j). \tag{5}$$

- 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 personalized lottery is the one associated to his type report instead of being chosen at random. In other words, the outcome is the lottery

$$\frac{1}{r}\left(\ell_i(t_i)+\sum_{j\in I^*\setminus\{i\}}\sum_{t_j\in T_j}\ell_j(t_j)dp_j(t_j)\right), \tag{6}$$

  where t_i is i's type report.

- In all other cases, ν^f coincides with μ^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 $K \geq 1$, ν^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 α^U denote the uniform

anchors (or, more generally, anchors that are atomless). We argue first that, for each individual i, $S_i^1(\nu^f|\alpha^U)$ is the set of reports $(\tau_i, 0)$ such that $\tau_i(t_i) \sim_i^f t_i$ 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

$$\frac{1}{r}\left(U_i(\ell_i(t_i)|t_i^*) + \sum_{j\in I^*\setminus\{i\}} U_i(\sum_{t_j\in T_j} \ell_j(t_j)dp_j(t_j)|t_i^*)\right),$$

while her expected utility under lottery (5) is equal to

$$\frac{1}{r}\left(U_i(\sum_{t_i\in T_i} \ell_i(t_i)dp(t_i)|t_i^*) + \sum_{j\in I^*\setminus\{i\}} U_i(\sum_{t_j\in T_j} \ell_j(t_j)dp_j(t_j)|t_i^*)\right).$$

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 ℓ_i. Remember also that this inequality is strict for all t_i such that $f(t_i, t_{-i}) \neq f(t_i^*, t_{-i})$ for some 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:

$$U_i(\ell_i(t_i^*)|t_i^*) > U_i(\sum_{t_i'\in T_i} \ell_i(t_i')dp_i(t_i')|t_i^*).$$

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 $t_i \sim_i^f t_i^*$.

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 ν^f coincides

with μ^f in the case relevant for computing $S_i^k(\nu^f|\alpha^U)$ for $k \geq 2$. $\square$

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 uniform beliefs (up to f-equivalent types). Once this is established, the social choice function f is used, as in μ^f, 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 $\sim_i^f$).

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 ν^f is not well defined in this case, as it would use f as well as the ℓ_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 ℓ_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' depths 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 conclusions of these two papers, in terms of the permissiveness of the results, are quite different, which should bring a word of caution towards our results, since we have restricted attention to single-valued rules.

3. Finally, assuming that behavior is governed by bounded levels of reasoning 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 by means of continuous mechanisms, 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

Abreu, D., and H. Matsushima (1992), "Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information," Unpublished Mimeo, Princeton University.

Arad, A., and A. Rubinstein (2012), "The 11-20 Money Request Game: A Level-k Reasoning Study," *American Economic Review* 102, 3561-3573.

Bergemann, D., S. Morris, and O. Tercieux (2011), "Rationalizable Implementation," *Journal of Economic Theory*, 146, 1253-1274.

Bosch-Domènech, A., J. García-Montalvo, R. Nagel, and A. Satorra (2002). "One, Two, (Three), Infinity, ...: Newspaper and Lab Beauty-Contest Experiments," *American Economic Review* 92, 1687-1701.

Cabrales, A., and R. Serrano (2011). "Implementation in Adaptative Better-Response Dynamics: Towards a General Theory of Bounded Rationality in Mechanisms," *Games and Economic Behavior* 73, 360-374.

Cai, H., and J. T.-Y. Wang (2006). "Overcommunication in Strategic Information Transmission Games," *Games and Economic Behavior* 56, 7-36.

Camerer, C., T.-H. Ho, and J.-K. Chong (2004), "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics* 119, 861-898.

Costa-Gomes, M., V. Crawford, and B. Broseta (2001). "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica* 69, 1193-1235.

Crawford, V. P. (2003). "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions." *American Economic Review* 93, 133-149.

Crawford, V. P. (2016). "Efficient Mechanisms for Level-k Bilateral Trading." Mimeo, Oxford University.

Crawford, V. P., and N. Iriberri (2007). "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica* 75, 1721-1770.

de Clippel, G. (2014), "Behavioral Implementation," *American Economic Review*, 104, 2975-3002.

de Clippel, G., R. Saran, and R. Serrano (2015), "Mechanism Design with Bounded Depth of Reasoning and Small Modeling Mistakes," Working Paper, Brown University.

Eliaz, K. (2002). "Fault-Tolerant Implementation," *Review of Economic Studies* 69, 589-610.

Glazer, J., and A. Rubinstein (2012). "A Model of Persuasion with a Boundedly Rational Agent," *Journal of Political Economy* 120, 1057-1082.

Glazer, J., and A. Rubinstein (2014). "Complex Questionnaires," *Econometrica* 82, 1529-1541.

Ho, T-H., C. Camerer, and K. Weigelt (1998). "Iterated Dominance and Iterated Best Response in Experimental "p-Beauty Contests," *American Economic Review* 88, 947-969.

Kunimoto, T. and R. Serrano (2016), "Rationalizable Implementation of Correspondences," Working Paper, Brown University.

Myerson, R. (1989), "Mechanism Design," in J. Eatwell, M. Milgate and P. Newman (eds.) *The New Palgrave: Allocation, Information, and Markets*, Norton, New York.

Myerson, R. B., and M. A. Satterthwaite (1983). "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory* 29, 265-281.

Nagel, R. (1995). "Unraveling in Guessing Games: An Experimental Study," *American Economic Review* 85, 1313-1326.

Oury, M., and O. Tercieux (2012), "Continuous Implementation," *Econometrica* 80, 1605-1637.

Saran, R. (2011). "Menu-Dependent Preferences and Revelation Principle," *Journal of Economic Theory* 146, 1712-1720.

Saran, R. (2016). "Bounded Depths of Rationality and Implementation with Complete Information," *Journal of Economic Theory* 165, 517-564.

Serrano, R., and R. Vohra (2005), "A Characterization of Virtual Bayesian Implementation," *Games and Economic Behavior* 50, 312-331.

Stahl, D. (1993), "Evolution of Smart-n individuals," *Games and Economic Behavior* 5, 604-617.

Stahl, D., and P. Wilson (1994), "Experimental Evidence on Individuals' Models of Other individuals," *Journal of Economic Behavior and Organization* 25, 309-327.

Stahl, D., and P. Wilson (1995), "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior* 10, 218-254.

Strzalecki, T. (2014), "Depth of Reasoning and Higher Order Beliefs," *Journal of Economic Behavior and Organization* 108, 108-122.

Wang, J. T.-Y., M. Spezio, and C. F. Camerer (2010). "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling

and Deception in Sender-Receiver Games," *American Economic Review* 100, 984-1007.