Level-k Mechanism Design[*]

Geoffroy de Clippel
Brown University

Rene Saran
Yale-NUS College

Roberto Serrano
Brown University

July 2016 Revised May 2017

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 weakening of standard strict incentive constraints. 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 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 Tilman Börgers, Sandro Brusco, Antonio Cabrales, Yan Chen, In-Koo Cho, Claude d'Aspremont, Olga Gorelkina, Paul Heidhues, Nagore Iriberri, Navin Kartik, Vijay Krishna, John Ledyard, Ludovic Renou, and Yuval Salant for comments, suggestions, and encouragement.

1 Introduction

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.[1]

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.[2] This paper is an attempt to understand the theoretical implications of level-k reasoning in mechanism design.

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 people's beliefs about how others would play the game instinctively, as a gut reaction without resorting to rational deliberation (Crawford, 2014).

The revelation principle (see, e.g., Myerson (1989) and the references therein)

[1] 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 ϵ-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, and de Clippel (2014) relaxes preference maximization. The discussion of the related literature below contains additional references.

[2] 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).

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?

Addressing all the aspects of this question is not feasible in a single paper. We are, therefore, somewhat limited in our scope. First, 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-$(k-1)$. This is for ease of exposition and without much loss of generality, as our results hold for a wide range of behavioral theories where participants see others as less sophisticated (see Remark 1 below).

Second, we investigate the implementability of *single-valued* social choice rules or *social choice functions.* These have the advantage of pinning down unequivocally the outcome that prevails, contingent on participants' information. This is the natural first step when studying implementation (see again, for example, Myerson (1989)).

Third, we want the desired outcomes to obtain for *all possible combinations* of positive depths of reasoning up to some level $K \geq 2$.[3] We opt for this robust approach to level-k implementation as depths of reasoning may vary from individual to individual, and even within a person, from mechanism to mechanism (e.g., Agranov *et al.* (2012) provide evidence that individuals depths of reasoning could vary depending on their expectation about the depths of others). The upper bound K is introduced to accommodate the experimental evidence suggesting that depths of reasoning are usually rather small. Interestingly, our results will be independent of the specific value of K.

Fourth, behavior with level-k reasoning can be highly sensitive to the level-0 anchor. Throughout the paper, we will thus discuss *many alternative scenarios regarding anchors.* This becomes especially relevant when discussing

[3] For reasons discussed later, our notion of implementability excludes the outcomes produced at level-0.

mechanism *design* beyond analyzing people's behavior in classic mechanisms such as a first-price auction or a double auction.

One could think at first that breaking the rational-expectations logic would lead to a very different set of implementable functions. In particular, one may conjecture that some functions can be implemented with level-k reasoning, but not in Nash equilibrium. This is not the case in our framework. Suppose the mechanism designer can pick the anchor in addition to picking the mechanism. The extent to which she can do this is debatable, but what matters for our point is that level-k implementation is *most permissive* under this scenario. 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 and shows the limits of the maximal set of implementable social choice functions. In fact, the restriction is stronger, as the incentive constrains must be satisfied with a strict inequality whenever the social choice function is responsive. We term this condition SIRBIC, which stands for "strict-if-responsive Bayesian incentive compatible" functions.[4]

As is often the case with versions of the revelation principle, the proof of Theorem 1 provides the main insight without involving intricate mathematical arguments. For an intuition, 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 of interest. 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 like in equilibrium, whenever $k > 1$, 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

[4] SIRBIC is a slight weakening of strict IC. For example, the optimal auction with a reserve price in Myerson (1981) satisfies SIRBIC but violates strict IC because of the types who do not participate in the auction; the same goes for the optimal bilateral trading mechanism in Myerson and Satterthwaite (1983) and the types who do not trade therein.

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).

Next, we show that the converse of Theorem 1 holds as well. This result reinforces Theorem 1 as it shows that level-k implementability imposes no additional restrictions beyond SIRBIC as far as the maximal set of implementable functions is concerned. Its applicability would be limited, though, if the anchors needed to achieve implementability were unreasonable. However, Theorem 2 is proved with truth-telling as an anchor in direct mechanisms: Truth-telling is often invoked as focal, as argued below.

Uniform anchors, in the sense of picking an action uniformly at random, are also often invoked in the literature, either to fit the behavior of experimental subjects in certain games, or more recently, when considering mechanism design (see Related Literature). In this context, one naturally wonders which social choice functions are level-k implementable with uniform anchors. We tackle this question in Section 6.

Our answer is two-fold. For independent private values, SIRBIC is sufficient once again when considering continuous social choice functions (Theorem 3). At the same time, an additional necessary condition (a kind of measurability condition) is identified for type-independent anchors in more general interdependent environments (Theorem 4). In many circumstances, this condition is sufficient once combined with SIRBIC (see online appendix). In our quest to draw conclusions that hold for a wide class of behavioral anchors, beyond the uniform case, the sufficiency results just mentioned also hold under arbitrary atomless anchors for mechanisms with a continuum of messages. The results are also robust to mixtures of truth-telling and atomless anchors. Section 6 ends with an important word of caution regarding sufficiency results for level-k mechanism design. We hope this discussion will foster further empirical and theoretical research on the topic.

Related Literature

The paper contributes to a recent literature at the intersection of mechanism design and behavioral economics (see references in footnote 1). Several papers have begun to investigate the implications of level-k behavior in classic mechanisms such as the first-price auction (Crawford and Iriberri (2007)), the double auction (Crawford (2016)) and the expected externality mechanism (Gorelkina (2017)).

Crawford *et al.* (2009) investigate how changing the reserve price in a first-price auction may affect the bidding behavior of two risk-neutral bidders, as a function of their depths of reasoning and their level-0 anchors. They provide numerical simulations and closed-form solutions for two classes of models with independent private values. Assuming that both bidders share the same anchor (truthful or uniform) and the same depth of reasoning (either depth one or depth two), they show that for some value distributions the optimal reserve price can be lower than Myerson's equilibrium reserve price, while it will be higher for other distributions. The seller can sometimes pick a reserve price that will give her a larger expected revenue than the maximal equilibrium profit, provided the likelihood of having two level-1 bidders with a uniform anchors is high enough. In that sense, Myerson's revenue equivalence result may break down. Finally, Crawford *et al.* design an 'exotic auction' that guarantees a possibly arbitrary high revenue when both bidders' depths of reasoning are odd, and a zero revenue otherwise. The concerns they express when discussing the realism of this mechanism and its potential for success share some common themes with the word of caution we give in Subsection 6.3.

Crawford (2016) pursues a related analysis in the case of bilateral trading with uniform anchors. He shows how Myerson and Satterthwaite's (1983) techniques to characterize incentive efficient mechanisms extend to the case of direct mechanisms where telling the truth is compatible with level-k behavior, provided the mechanism designer knows the depth of reasoning of both the buyer and the seller. Without that knowledge, getting truth-telling for all depths of reasoning requires using a random posted price. However, Crawford also shows how mechanisms that guarantee truth-telling are unduly restrictive, and discusses the relative performance of double auctions with reserve prices in some examples with uniform values.

The present paper contrasts with this body of work in multiple ways. First, we tackle the question of mechanism design in a *general framework*, that includes the problems discussed above as particular cases. Second, we consider *larger classes of level-*0 *anchors*. Third, following the tradition of implementation theory, we look for mechanisms that deliver the *right outcome independently of the combination of depths of reasoning.* Papers by Crawford and co-authors elegantly illustrate how social choice functions that are not Bayesian incentive compatible can be level-k implementable in auctions and simple trading settings but, as we argue later, achieving this requires either (a) the assumption that all players are homogeneous in their depths of reasoning (see Example 1) or (b) implementing different social choice functions at different depths (as in the 'exotic auction' of Crawford *et al.* mentioned above and Example 1). While the assumption in (a) is questionable on grounds of robustness, implementing different social choice functions at different depths suggests a new direction for mechanism design theory. But we ought to tread that path with caution because discrimination based on depths of reasoning lacks normative justification (since depths do not determine preferences) and may not be optimal without precise knowledge of the distribution of depths of reasoning. In contrast, our results apply independently of the pool of (inexperienced) subjects the mechanism designer faces. Conclusions one can reach regarding achievable profits, for instance, do not hold in expectation over possible depths of reasoning, but hold instead regardless of their distribution. This prevents incurring potentially serious losses, or missing one's social goal, from holding incorrect beliefs regarding participants' depths of reasoning.

While this paper shows how bounded depths of reasoning and equilibrium logic can entail remarkably similar restrictions on social choice functions when it comes to their implementability, they can also vary greatly in other dimensions. If one insists on robustness to small modeling mistakes, for instance, Bayesian Nash implementation becomes very restrictive (requiring a strong form of Maskin monotonicity beyond incentive compatibility; see Oury and Tercieux (2012)). However, as shown in a companion paper (see de Clippel *et* *al.* (2015)), continuous implementation with bounded levels of reasoning relies only on the continuity of the social choice function beyond SIRBIC.

Our notion of implementation also shares some commonality with rationalizable full implementation, in particular the iterative construction rooted in best responses that can accommodate a wide variety of reasonings and behaviors. It is less demanding, though, as individuals' depth of reasoning is bounded and behavior at cognitive state of depth 0 is fixed. Bergemann *et al.* (2011) study 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 next, and to figure out in particular whether implementation can be significantly more permissive when behavior is better described via level-k reasoning than via Bayesian Nash equilibrium (see Example 2 below for an illustration). In a related paper concurrent to de Clippel *et al.* (2015), Saran (2016) investigates the impact that an upper-bound on depths of reasoning has on the rationalizable implementation of single-valued social choice functions, this time under complete information. In this case, the desired outcome must obtain for all anchors, which is naturally very demanding (requiring for instance strategy-proofness and 'strong non-bossiness' in the case of an 'independent domain of preferences', along with a strong notion of monotonicity, more generally).

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.[5] 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* $p(\cdot|t_i)$. We assume throughout the paper that the *marginal probability distribution* p_i over T_i has full support for all individuals i. This assumption is made only for notational convenience, as results extend otherwise by dropping types with zero probability. 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 (of course, a player's strategy will depend on her depth of reasoning, as discussed below). A strategy profile σ and type profile t induce a lottery $\mu(\sigma(t))$ over X.[6]

For any social choice function f, say that an individual i is *irrelevant for* f if $f(t_i, t_{-i}) = f(t'_i, t_{-i})$, for all $t_i, t'_i \in T_i$ and all $t_{-i} \in T_{-i}$. Thus i's type matters under no circumstance when i is irrelevant. Individuals who are not irrelevant are called *relevant.* Social choice functions in this paper are assumed to treat all individuals as relevant. This is for notational convenience only, as all results extend to the problems with irrelevant individuals as well, simply by having the mechanism designer overlook their reports in mechanisms.

[5] We 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.

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

We conclude the section with several technical observations. Throughout the paper, it is assumed that the sets and functions considered have the right structure to ensure 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 paper for the level-k model. In the remark below, 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* 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 $\alpha = (\alpha_i)_{i \in I}$. 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 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 $k \geq 1$, 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 $\sigma_{-i} \in S^k_{-i}(\mu|\alpha)$. The index k is called an individual's *depth of reasoning*.

Remark 1. We present 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. Using simple induction arguments, all our results can easily be adapted to a wide class of theories where individuals see others as less sophisticated than themselves. This would include, for instance, all the theories described through the language of cognitive hierarchies (Strzalecki (2014)), which subsumes earlier models by Stahl (1993), Stahl and Wilson (1994, 1995), and Camerer *et al.* (2004) among others. $\diamond$

It has been argued that, for many subjects in the lab, their depths of reasoning are probably rather small. At the same time, such depths vary from individual to individual, and, even within a person, they may vary from mechanism to mechanism. It is currently not well understood how one could identify or impact individuals' depth of reasoning. Therefore, we do not fix the designer's beliefs about the distribution of depths of reasoning among the individuals. Instead, we introduce an upper bound $K \geq 2$ on the individuals' depths of reasoning, and assume that the mechanism designer thinks that all combinations of depths between 1 and K are in principle possible. Our results are robust to specific assumptions regarding the distribution of depths, as long as the upper bound K is larger or equal to 2 and the distribution assigns positive probability on all profiles of depths in $\{1, \ldots, K\}^I$.[7] 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.

[7] In fact, even weaker assumptions on the distribution of depths suffice for our results. For instance, the general necessary condition in Theorem 1 holds as long as for each individual i, the distribution of depths 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 $j \neq i$.

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. 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' beliefs regarding others' gut feelings towards the mechanism, and hence, does not see herself as trying to affect those.[8] 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

[8] If level-0 is not just a belief about others but, in fact, corresponds to actual unsophisticated behavior, then until we better understand how to affect it, we have to resign to the fact that actions at level-0 will in general not generate the social choice outcome.

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 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 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 $t_i \sim_i^f t'_i$, 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,

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

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 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 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 $\sigma_i^2(t_i')$.[9] 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. □

The reader is referred to the Introduction for an intuitive outline of the main part of the previous proof. We next present two examples, which are

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

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 usual 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

$$\begin{array}{cl} \frac{2}{3}v+\frac{1}{9}, & \text{if } v \geq \frac{1}{3} \\ v, & \text{if } v < \frac{1}{3} \end{array}.$$

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

$$\begin{array}{cl} \frac{2}{3}c+\frac{2}{9}, & \text{if } c \leq \frac{2}{3} \\ c, & \text{if } c > \frac{2}{3}. \end{array}$$

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. This observation is consistent with our earlier claim that Bayesian incentive compatibility is not necessary when players' levels are homogeneous. But now, the buyer's (seller's) level-2 consistent strategy when paired with the seller's (buyer's) level-1 consistent strategy results in an outcome that is not incentive compatible either. So how are we able to get around incentive compatibility even though we have heterogeneous levels?

The key is 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.[10] While such a differential treatment might be questionable on normative grounds (because levels of reasoning do not determine individual preferences) and may not be optimal without specific knowledge of the distribution of levels, the example clearly suggests a new direction for the theory,

[10] 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.

in which an expanded notion of a type may include cognitive sophistication, thereby making the social choice function contingent on such considerations.$\diamond$

The second example has implications that may take us far afield from the current paper. It shows that Bayesian incentive compatibility ceases to be necessary if the designer wants to implement a social choice set.[11]

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 μ: 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 α 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 ones 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 $k \geq 1$.

[11] Note that a social choice set instead of a social choice correspondence is the appropriate notion of set-valued rules in case of incomplete information.

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. $\diamond$

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 our next results show 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.[12] 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.

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

Lemma 1. *Let f be a social choice function. Then $f(t) = f(t')$ 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^* 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 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 $S_i^1(f|\alpha^*)$. 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 $\sigma_{-i}^1 \in S_{-i}^1(f|\alpha^*)$. Her expected utility from reporting type t_i' when of type t_i is

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

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

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 the level-0 anchor in the direct mechanism is the uniform distribution). 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 Uniform Anchors and Beyond in General Mechanisms

Anchors α in a mechanism $\mu : M_1 \times \cdots \times M_I \to \Delta X$ are *uniform* if α_i is the uniform probability distribution over M_i, for each individual i. While often used in the literature on level-k reasoning, they are peculiar in several respects. For instance, uniform anchors have full support. By contrast, it is plausible that, even without rational deliberation, bidders would not bid below the reserve price when they value the good more. Uniform anchors also are type-independent. Instead, an individual's gut reaction may be biased towards the truth.

Fortunately, the sufficiency results we derive in this section are robust to the extent that they hold for a wide class of anchors beyond the uniform case (see, however, the word of caution in Subsection 6.3). 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. To be clear, we will show how, for most SIRBIC social choice functions, there is often a mechanism that level-k implements it for *all* profiles of atomless anchors. As we will see, that same mechanism will also succeed should anchors be truthful instead.

6.1 Independent Private Values

Individuals have *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.

Consider now the following mechanism μ^f. Each individual reports a type along with a real number between -1 and 1. Say $m_i = (t_i, z_i) \in T_i \times [-1, 1]$, for each i. Based on these reports, the designer applies f to a profile of types selected as follows: for each individual i, if $z_i = kz_j$ for all $j \neq i$ and some integer k, then the designer uses the reported t_i; otherwise, she uses a type picked at random according to the density p_i.

Theorem 3. *Consider an environment with independent private values. If a continuous social choice function f satisfies SIRBIC, then for all $K \geq 1$, μ^f implements f up to level-K given atomless anchors.*

Proof. Fix atomless anchors α. We claim that $S_i^1(\mu^f|\alpha)$ is the set of strategies for which individual i of type t_i sends a message $(\tau_i, 0)$ such that $\tau_i \sim_i^f t_i$. Pick the level-1 individual i. She believes that all $j \neq i$ are level-0 players playing an atomless strategy α_j. Therefore, the probability that the realized value of z_j is equal to kz_i for any z_i is equal to zero. Hence, from the point of view of the level-1 individual i, the mechanism designer will almost surely use for j a type picked at random according to the density p_j. In addition, the level-1 individual i expects with probability 1 to have the mechanism designer overlook his type report and instead use a type drawn according to p_i when reporting a nonzero z_i (this is because the level-0 of all $j \neq i$ play atomless strategies, and so for any nonzero z_i, the probability that the realized value of z_j is such that $z_i = kz_j$ is equal to zero). In contrast, reporting $z_i = 0$ guarantees that the mechanism will use i's reported t_i because then $z_i = 0 \times z_j$ for all $j \neq i$ and z_j. To summarize, i expects the lottery

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

when sending a message with a nonzero z_i, and expects the lottery

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

when sending a message $(t_i, 0)$.

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 and there exists t_i such that $t_i \not\sim_i^f t_i^*$ (because individual i is relevant for f), there is a positive p_i-measure of types t_i 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 nonzero number is never a best response against atomless 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 $S_i^1(\mu^f|\alpha)$ is the set of strategies such that i of type t_i picks a message $(\tau_i, 0)$ with $\tau_i \sim_i^f t_i$.

We now show that $S_i^k(\mu^f|\alpha) = S_i^1(\mu^f|\alpha)$, 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 atomless anchors, thanks to Lemma 1. Level-2 of individual i believes that level-1 of any individual $j \neq i$ plays according to strategies in $S_j^1(\mu^f|\alpha)$. Since level-1 of j sends a zero along with her type report, the designer will accept j's type report. 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 $\sim_i^f$). Since level-2 of individual i believes that all others are reporting zero, she expects the lottery (2) if she sends a nonzero 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 the level-1 of individual i, 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 $S_i^2(\mu^f|\alpha) = S_i^1(\mu^f|\alpha)$. The argument extends trivially to any higher depth of reasoning $k > 2$. □

The mechanism μ^f succeeds by effectively separating individuals' beliefs when having a depth of reasoning 1 or 2+. Under atomless anchors, a level-1 individual expects that the mechanism designer overlooks others' type reports. By using instead arbitrary types drawn according to the true empirical distribution of types, this individual faces the same expected outcome under f as if others where truth-telling.[13] A level 2+ individual expects that others will submit a zero, in which case the mechanism designer does take reported types into account, but also that type reports are truthful.

Remark 2. In many classic implementation problems, including auction and bilateral trade problems (see Crawford and Iriberri (2007), Crawford *et al.* (2009), and Crawford (2016)), type sets are intervals. In such cases, any continuous SIRBIC social choice function f can be level-k implemented given atomless anchors by a *direct* mechanism (having $M_i = T_i$). To construct a mechanism achieving this, observe that for each i there exists a bijection b_i between T_i and a subset $A_i \subseteq T_i$ of measure zero. An analogue to Theorem 3 holds when using the mechanism $\hat{\mu}^f$ obtained by applying f to types selected as follows: for each individual i, if $t_i \in A_i$, then the designer uses the type $b_i^{-1}(t_i)$; otherwise, she uses a type picked at random according to the density p_i.

Remark 3. In mechanism μ^f, we do not need the level-0 anchors to be atom-

[13] In contrast, using f as a direct mechanism would typically fail at this stage since p_i need not match the anchor α_i.

less over the type space. In particular, suppose the level-0 reports her type truthfully but picks a number in $[-1,1]$ according to some atomless distribution. It is easy to check that μ^f also level-k implements f with such anchors that are truthful over the type space. Similarly, in the context of the previous remark, the direct mechanism $\hat{\mu}^f$ also level-k implements f with truthful anchors.

6.2 General Interdependent Values

Beyond the case of independent private values, level-k implementation given uniform anchors entails a restriction in addition to SIRBIC. The next example provides a first intuition.

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:

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

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

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. In fact, the implementability issue in this example does not only arise with uniform anchors, but instead *as soon as anchors are type-independent.*

Individual i's (interim) preference over state-independent or constant lotteries, i.e., over $\Delta(X)$, when of type t_i is represented by the following utility function:

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

The next condition requires the social choice function to be responsive to t_i versus t'_i only if the two types define different preferences.

Definition 2. The social choice function f is *responsive only when preferences differ* if $t_i \not\sim_i^f t'_i$ implies that individual i has different interim preferences over constant lotteries at t_i, t'_i, that is, there do not exist $\lambda > 0$ and β such that $U_i(\cdot|t_i) = \lambda U_i(\cdot|t'_i) + \beta$.

Remark 4. This 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.

When the mechanism designer is bound to type-independent (e.g., uniform) anchors, implementability requires the social choice function to be responsive only when preferences differ in addition to SIRBIC.[14]

Theorem 4. *If a social choice function f is implementable up to level-K given type-independent anchors, then f is responsive only when preferences differ.*

Proof. Let μ be a mechanism that implements f up to level-K given type independent anchors α. For each individual i, let σ_i^1 be some level-1 consistent strategy, that is, $\sigma_i^1 \in S_i^1(\mu|\alpha)$. 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,

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

[14] This result does not contradict the sufficiency result in Theorem 3. Indeed, it is not difficult to check that under independent private values, any social choice function that satisfies SIRBIC must be responsive only when preferences differ. Example 3 shows how this implication does not hold more generally.

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 that level-1 individual i 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 $S_i^1(\mu|\alpha)$. 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 responsive only when preferences differ. □

Returning to Example 3, note how both types of each agent have identical interim preferences over constant lotteries. Thus, being responsive only when preferences differ 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.

Does level-k implementation with type-independent entail further restrictions beyond SIRBIC and beyond the responsiveness only when preferences differ? In a supplementary appendix available online, we answer this question in the negative. That is, we extend the technique used to prove Theorem 3, and show that these two properties are sufficient in most problems.[15]

6.3 A Word of Caution

Assuming that anchors are uniform makes sense if individuals think that others' gut reaction to a game would be totally random. This is plausible, for instance, if level-1 players see others as not paying attention to, or not understanding the rules of the mechanism. However, other anchors may be sensible too.

It is preferable then to obtain sufficiency results that remain valid for a wide set of anchors.[16] We derived results, for instance, that remain valid

[15] The online appendix is available at https://goo.gl/Dse9AK.

[16] Social choice functions that are implementable in strictly dominant strategies will be level-k implementable for all anchors. An issue, of course, is that few social choice functions may be implementable in that sense.

for any combination of atomless anchors (as well as truth-telling). While we restricted attention to problems where any two individuals share the same anchors regarding third parties, our sufficiency results easily extend to cases with personalized anchors as well.

That being said, it is entirely possible that reporting zero (along with a type) when participating in μ^f is salient enough that anchors would display an atom at zero. But, of course, when f is implementable, μ^f is not the only mechanism implementing it with atomless anchors. For instance, the proof of Theorem 3 can be adapted to show that the following alternative mechanism would work too: messages remain unchanged, but the mechanism designer uses an individual i's actual type report if and only if the real number he sent along falls in a given finite subset A of $[-1, 1]$; otherwise she uses an arbitrary type picked according to p_i. In the spirit of framing effects, the set A can be presented in different ways, e.g. as a list or as the set of solutions to some equation.[17]

More generally, we conjecture that level-0 anchors might be *mechanism-specific*. We see this paper as setting a benchmark for future advances. Progress on this topic will likely come from the combination of empirical and theoretical work. Mechanisms derived theoretically, such as those in this paper for a start, should be tested empirically, and empirical lessons as to how anchors may vary with the game being played and its description should inform theorists when designing new mechanisms. Theorem 1 remains applicable, however, and thus one should always keep in mind the restrictions imposed by SIRBIC.

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.

[17] 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.

Agranov, M., E. Potamites, A. Schotter, and C. Tergiman (2012), "Beliefs and Endogenous Cognitive Levels: An Experimental Study," *Games and Economic Behavior* 75, 449-463.

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. (2014). "A Comment on How Portable is Level-0 Behavior? A Test of Level-k Theory in Games with Non-neutral Frames by Heap, Rojo-Arjona, and Sugden." Mimeo, Oxford University and UCSD.

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

Crawford, V. P., R. Kubler, Z. Neeman, and A. Pauzner (2009). "Behaviorally Optimal Auction Design: Examples and Observations," *Journal of the European Economic Association* 7, 377-387.

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.

Gorelkina, O. (2015), "The Expected Externality Mechanism in a Level-k Environment," Working Paper, Max Planck Institute, Bonn.

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. B. (1981), "Optimal Auction Design," *Mathematics of Operations Research* 6, 58-73.

Myerson, R. B. (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.

Renou, L. and K. H. Schlag (2011), "Implementation in Minimax Regret Equilibrium," *Games and Economic Behavior* 71, 527-533.

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.

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.