BROWN
Orlando Bravo Center
for Economic Research

Continuous Level-*k* Mechanism Design[*]

Bravo Working Paper # 2021-002

Geoffroy de Clippel[†] Rene Saran[‡] Roberto Serrano[§]

Abstract: In de Clippel, Saran, and Serrano (2019), it is shown that, perhaps surprisingly, the set of implementable social choice functions is essentially the same whether agents have bounded depth of reasoning or rational expectations. The picture is quite different when taking into account the possibility of small modeling mistakes. While continuous strict implementation becomes very demanding (Oury and Tercieux (2012) – OT), continuity in level-*k* implementation obtains essentially for free. A decomposition of the conditions implied by the OT implementation notion confirms that it is the use of equilibrium, and not continuity per se, that is responsible for the difference.

JEL Classification: C72, D70, D78, D82.
Keywords: mechanism design; bounded rationality; level-*k* reasoning; small modeling mistakes; incentive compatibility; continuity.

1 Introduction

Building institutions that are resilient to mispecifications of basic assumptions is an important task for economists. In the mechanism design literature, concerned with

[*] We thank Vince Crawford, Jack Fanning, Amanda Friedenberg, Willemien Kets, Laurent Mathevet, Stephen Morris, Arunava Sen, Olivier Tercieux, and audiences at NYU, ISI (Delhi), Singapore

[†] Department of Economics, Brown University, declippel@brown.edu.
[‡] Department of Economics, University of Cincinnati, saran.rene@gmail.com
[§] Department of Economics, Brown University, roberto_serrano@brown.edu.

Continuous Level-k Mechanism Design[*]

Geoffroy de Clippel[†] Rene Saran[‡] Roberto Serrano[§]

This version: Feb 2021

Abstract

In de Clippel, Saran, and Serrano (2019), it is shown that, perhaps surprisingly, the set of implementable social choice functions is essentially the same whether agents have bounded depth of reasoning or rational expectations. The picture is quite different when taking into account the possibility of small modeling mistakes. While continuous strict implementation becomes very demanding (Oury and Tercieux (2012) – OT), continuity in level-k implementation obtains essentially for free. A decomposition of the conditions implied by the OT implementation notion confirms that it is the use of equilibrium, and not continuity *per se*, that is responsible for the difference.

JEL Classification: C72, D70, D78, D82.

Keywords: mechanism design; bounded rationality; level-k reasoning; small modeling mistakes; incentive compatibility; continuity.

1 Introduction

Building institutions that are resilient to mispecifications of basic assumptions is an important task for economists. In the mechanism design literature, concerned with

[*] We thank Vince Crawford, Jack Fanning, Amanda Friedenberg, Willemien Kets, Laurent Mathevet, Stephen Morris, Arunava Sen, Olivier Tercieux, and audiences at NYU, ISI (Delhi), Singapore Economic Theory Workshop (2014), SCW 2014 (Boston), and SAET 2015 (Cambridge) for their comments and suggestions on previous drafts.

[†] Department of Economics, Brown University, declippel@brown.edu

[‡] Department of Economics, University of Cincinnati, saran.rene@gmail.com

[§] Department of Economics, Brown University, roberto_serrano@brown.edu

the exploration of institutions in which the informational constraints of the designer are incorporated, such resilience or robustness has been addressed in several ways. The approach to robustness followed in the current study relies on a local analysis. The model is tested against small mistakes in the assumptions. From this point of view, our paper continues the methodology employed in Oury and Tercieux (2012, henceforth OT) and Jehiel *et al.* (2012).[1]

But what sets this paper aside from all the work mentioned so far lies in how individual behavior is modeled, considering participants with bounded depth of reasoning.[2] As it turns out, the exact size of that bound will be of no significance for our results. Rather, what will matter is the existence of such a bound, whatever it is, which will render our conclusions markedly different from those based on equilibrium analysis.

To capture these ideas, we propose to study OT's notion of continuous implementation for the theory of level-k mechanism design introduced in de Clippel, Saran and Serrano (2019, henceforth dCSS).[3] To make results as transparent and accessible as possible, we restrict attention in most of the paper to a simpler framework with (i) simple type spaces as the benchmark model for the planner (where the entire belief hierarchy is determined by the first-order beliefs), (ii) simple mechanisms (individuals report the first-order beliefs), and (iii) truthful behavioral anchors (using truth-telling as level-0). However, none of these expositional assumptions play a crucial role in our results. Extensions are discussed in the final section, and presented in greater detail

[1] Other robustness checks with respect to information structures in mechanism design include Chung and Ely (2003) for undominated Nash implementation, Aghion *et al.* (2012) for subgame-perfect implementation, and Neeman (2004) and Heifetz and Neeman (2006) in the full surplus extraction problem. See also McLean and Postlewaite (2002) and Weinstein and Yildiz (2007) for related robustness concerns beyond implementation. For other approaches that model versions of global (as opposed to local) robustness, in the sense that the model is tested against a wide class of mispecifications, see, e.g., Bergemann and Morris (2005, 2012), Artemov *et al.* (2013), Ollár and Penta (2017), and Lopomo *et al.* (2020).

[2] Bounded depth of reasoning can provide a better description of behavior than equilibrium models, especially when participants are inexperienced. For evidence of this in various contexts, see for instance Rapoport and Amaldoss (2000), Costa-Gomes *et al.* (2001), and Katok *et al.* (2002) for iterated elimination of strictly dominated strategies; Nagel (1995), Ho *et al.* (1998), and Bosch-Domènech *et al.* (2002) for iterated elimination of weakly dominated strategies; and Binmore *et al.* (2002) for backward induction.

[3] For other recent applications of level-k theory to mechanism design, see Kneeland (2020) and Crawford (forthcoming).

in an Online Appendix.

In a nutshell, the main result of the paper is this: any continuous social choice function (SCF) that is level-k implementable for k smaller than a fixed, arbitrary upper bound K is also continuously implementable in this sense. The proof is not obvious because, contrary to what one may conjecture at first, the level-k correspondence can fail to be upper hemicontinuous with respect to the players' information (see Section 2). A main issue is that there may be sequences of types under which a level-k player ($k \geq 2$) views others' actions as correlated with the state, while no such correlation is possible at the limit type. In that case, mechanisms implementing the desired SCF over the planner's model will have to be continuously extended with care; see Section 4.

Independently of the difficulty of its proof, we find the result surprising and noteworthy. Indeed, dCSS shows that an SCF is level-k implementable if and only if it satisfies SIRBIC,[4] a mild strengthening of Bayesian incentive compatibility. SIRBIC is also equivalent to weak Bayesian implementation in strict equilibrium. But things diverge dramatically when adding robustness to small modeling mistakes. While considerably reducing the set of achievable SCFs under the equilibrium paradigm (see OT), continuous level-k implementation obtains almost for free under SIRBIC.

As a step to elucidate this difference, we pursue the intuition of the revelation principle in Section 5 to show that continuous strict implementation in OT's sense is possible only if the SCF admits for each larger type space a continuous extension that (a) satisfies SIRBIC with respect to types in the planner's model, and (b) is Bayesian incentive compatible. Phrased in terms of SCFs, with no explicit reference to Bayesian Nash equilibria of mechanisms, this necessary condition facilitates comparisons with our notion of continuous implementation. The latter is indeed equivalent to the modified condition where (b) is simply dropped. We close Section 5 by illustrating how demanding (b) can be, first by means of a transparent example and then by proving that our necessary condition for continuous strict equilibrium implementation implies strict interim rationalizable monotonicity, a very restrictive condition that, for example, boils down to a strengthening of Maskin monotonicity (Maskin (1999)) in complete-information environments. We remark that no such

[4] SIRBIC stands for 'Strict-if-Responsive Bayesian Incentive Compatibility.'

connection to monotonicity conditions is found for the level-k implementation model.

The plan of the paper is as follows. Section 2 presents an important example explaining the lack of upper hemicontinuity of level-k behavior, a point that is of relevance for general Bayesian games. Section 3 describes the elements of our model. Section 4 provides our main characterization result of continuous level-k implementation in terms of SIRBIC (Theorem 1). Section 5 discusses at length the connections between our approach and that in OT, including a decomposition of the implications of their implementation notion, which lead to our Theorems 2 and 3, pinning down exactly the source of the difference in the results of the two approaches. Section 6 concludes with a brief discussion of generalizations and applications. Some proofs are relegated to an Appendix, and an Online Appendix presents more details on generalizations and applications of our approach.[5]

2 Level-k Fails Upper Hemicontinuity: An Example

In this preliminary section, we make a point that is relevant for general Bayesian games. Following Harsanyi (1967, 1968), individuals' information is described by a *state space* Θ and a *type space* $\mathcal{T} = (T_i, \pi_i)_{i \in I}$, where T_i is the set of types of individual i, assumed to be a compact metric space, and $\pi_i : T_i \to \Delta(\Theta \times T_{-i})$ is a continuous function, specifying the beliefs $\pi_i(t_i)$ over the realized state of the world and other individuals' types for each type t_i of individual i.[6] Player i's payoff is a continuous function of the state and players' chosen actions.

It follows from Berge's Maximum Theorem (Berge (1963)) that best-response correspondences are upper hemicontinuous: Fix a player i in a Bayesian game with a compact action space for each player, and a continuous strategy σ_j for each $j \neq i$. Take a sequence $(t_i^n)_{n \geq 1}$ of types converging to some t_i^*, and a sequence $a_i(t_i^n)$ of best

[5] The Online Appendix is available at `https://tinyurl.com/1wxolg5z`.

[6] We keep the standard definition of a type space in order to have a clear comparison with OT. Our point of departure from OT lies in the modeling of individual behavior, equilibrium for OT and level-k for us. See Kets (2017) and Heifetz and Kets (2018) for the construction of type spaces when departing from the assumption that players have infinite ability to reason about each other's beliefs. These papers show how equilibrium and rationalizability predictions are sensitive to such departures.

responses against σ_{-i}. If $a_i(t_i^n) \to a_i^*$, then the limit action a_i^* is a best response against σ_{-i} for player i of type t_i^*.

Since level-k behavior is associated with iterations of best-responses, one might expect level-k behavior itself to be upper hemicontinuous with respect to types when anchors are continuous. But this intuition is flawed, as the following example shows.

Example 1. Consider the following two-player Bayesian game. There are two states of the world, θ and θ'. Players' types and beliefs are defined as follows: $T_i = S_i \times Z_i$, where $S_i = \{0, 0.4, 0.8\}$ and $Z_i = [0, 1]$, for all $i \in \{1, 2\}$. Player i of type $t_i = (s_i, z_i)$ places probability $(s_i + 0.2z_i)/3$ on $\big(\theta, (s_j, z_i)\big)$ and probability $(1 - s_i - 0.2z_i)/3$ on $\big(\theta', (s_j, 1 - z_i)\big)$ for all $s_j \in S_j$. Notice that this is a simple type space, i.e., distinct types have distinct first-order beliefs.

The two players play a direct mechanism in which player i is asked to report his type $(\hat{s}_i, \hat{z}_i) \in T_i$. The outcome, however, is determined solely on the basis of the first components of the reports, $\hat{s}_1$ and $\hat{s}_2$. The associated payoffs are defined in the following tables:

$\hat{s}_1$ \ $\hat{s}_2$	0	0.4	0.8
0	$(0, 0)$	$(0, 0)$	$(0, 1)$
0.4	$(0, 0)$	$(0, 0)$	$(0, -1.5)$
0.8	$(1, 0)$	$(-1.5, 0)$	$(-1, -1)$

State θ

$\hat{s}_1$ \ $\hat{s}_2$	0	0.4	0.8
0	$(0, 0)$	$(0, 0)$	$(0, -1.5)$
0.4	$(0, 0)$	$(0, 0)$	$(0, 1)$
0.8	$(-1.5, 0)$	$(1, 0)$	$(-1, -1)$

State θ'

Table 1: Payoff Functions.

The reader is referred to Section 3 below for detailed definitions of concepts in the level-k model, which we proceed to illustrate here. We assume truthful anchors, meaning that a level-1 player assumes his opponent will report his type truthfully. In that case, regardless of his type, the level-1 of player i believes that, in each state, his opponent is equally likely to report any $\hat{s}_j \in S_j$. The expected payoffs are then state-independent, and any report with $\hat{s}_i$ equal to either 0 or 0.4 is a best response whatever player i's type (giving a zero expected payoff in both states, whereas any report with $\hat{s}_i = 0.8$ gives an expected payoff of -0.5 in both states).

In particular, the following strategy

$$\sigma_2(t_2) = \begin{cases} (0,0), & \text{for all } t_2 = (s_2, z_2) \text{ s.t. } z_2 \leq 1/2 \\ (0.4,0), & \text{for all } t_2 = (s_2, z_2) \text{ s.t. } z_2 > 1/2 \end{cases}$$

is a level-1 strategy for player 2 (i.e., a best response to the assumed truthful behavioral anchor).

For level-2 reasoning, we identify best responses to the belief that the opponent is a level-1 agent. In particular, against the above strategy of player 2, any report with $\hat{s}_1 = 0.8$ is a best response for all types $t_1 = (0.4, z_1)$ of player 1 such that $z_1 < 1/2$. Indeed, in that case, such a type of player 1 believes that player 2 reports $(0,0)$ when the state is θ (with probability $0.4 + 0.2z_1$) and $(0.4, 0)$ when the state is θ' (with probability $0.6 - 0.2z_1$). But the best response against σ_2 is such that $\hat{s}_1$ is equal to either 0 or 0.4 when $t_1 = (0.4, 1/2)$. Berge's theorem fails to apply because σ_2 is discontinuous.

Now, could it be that a report with $\hat{s}_1 = 0.8$ is a best-response for player 1 of level-2 and type $t_1 = (0.4, 1/2)$ against another level-1 strategy for player 2? The answer is "No", and here is the reason. Since the mechanism ignores the second component of player 2's report, what is important for player 1 is the marginal distribution of $\hat{s}_2$ in each state. As argued above, any level-1 strategy for player 2 must be such that the marginal distribution of $\hat{s}_2$ in each state puts zero probability on $\hat{s}_2 = 0.8$. Player 1 of type $t_1 = (0.4, 1/2)$ believes that both states are equally likely and, in each state, player 2 is also equally likely to be of type $(0, 1/2)$, $(0.4, 1/2)$ or $(0.8, 1/2)$. Hence, from the perspective of type $t_1 = (0.4, 1/2)$, the marginal distribution of $\hat{s}_2$ generated by any level-1 strategy for player 2 is *independent of the state* and puts a positive probability on either 0 or 0.4. Then the best response for player 1 of type $t_1 = (0.4, 1/2)$ is to submit a report such that $\hat{s}_1$ is either equal to 0 or 0.4, but not 0.8.

Notice that, as long as $t_1 = (0.4, z_1)$ is such that $z_1 < 1/2$, player 1 believes that 2's action is correlated with the states when 2 employs σ_2, but such a correlation is lost when $t_1 = (0.4, 1/2)$. This is the culprit for the level-2 correspondence to fail upper hemicontinuity.

The preceding example highlights the difficulties entailed in the model of level-k reasoning when it comes to issues of continuity. Nevertheless, the current paper poses the question of what rules can be implementable if one insists on the continuity *desideratum* under level-k reasoning, and in doing so, allows a comparison with the model of (fully rational) equilibrium implementation. We turn to it next.

3 The Model

Our setting essentially mirrors that of OT, although we do not require their assumption that type spaces be countable for our first two results (Theorems 1 and 2). To enhance the comparability with their main conclusions, we shall assume countability for Theorem 3.

Alternatives, States, and Utility Functions: A social planner/mechanism designer needs to select an alternative from a set X, which is assumed to be a compact metric space. Her decision impacts the satisfaction of individuals in a finite set I. Unfortunately, she does not know their preferences. Formally, individual i's *preference* is represented by a continuous and bounded Bernoulli function $u_i : X \times \Theta \to \mathbb{R}$, where Θ is the set of states, assumed to be a compact metric space. Individual i evaluates any lottery $\ell \in \Delta X$ by its expected utility $U_i(\ell, \theta) = \int_{x \in X} u_i(x, \theta) d\ell$.

Belief Hierarchies: Recall the definition of type spaces at the start of Section 2. Given a type space $\mathcal{T}$ and individual i, let $q_i^1 : T_i \to \Delta\Theta$ be a function such that $q_i^1(t_i)$ is the *first-order belief* of type t_i, i.e., the belief about the realized state, which is equal to the marginal distribution of $\pi_i(t_i)$ on Θ. We can further describe the *second-order belief* $q_i^2(t_i)$ of type t_i (i.e., a belief about the realized state and the other individuals' first-order beliefs) as follows:

$$q_i^2(t_i)(E) = \pi_i(t_i)(\{(\theta, t_{-i}) : (\theta, (q_j^1(t_j))_{j \neq i}) \in E\}),$$

for all measurable $E \subseteq \Theta \times (\Delta\Theta)^{I-1}$. Continuing in this manner, we can describe the z^{th}*-order belief* $q_i^z(t_i)$ of type t_i, which specifies t_i's belief regarding the realized state and up to $(z-1)$ orders of beliefs of the other individuals. Thus, we can associate an infinite hierarchy of beliefs $q_i(t) = (q_i^1(t_i), q_i^2(t_i), \ldots)$ to any type t_i in any type space $\mathcal{T}$. The belief hierarchy $q_i(t_i)$ is coherent in the sense that beliefs at different orders do not contradict each other. For each $z \geq 1$, let $Q_i^z(T)$ denote the range of q_i^z, that is, $Q_i^z(T) = \{q_i^z(t_i) : t_i \in T_i\}$.

The collection of all infinite hierarchies of beliefs for which it is common knowledge that the beliefs are coherent defines a type space $\mathcal{T}^* = (T_i^*, \pi_i^*)_{i\in I}$, which is the *universal type space* generated by Θ (see Mertens and Zamir (1985), Brandenburger and Dekel (1993)). Recall that, under the product topology, T_i^* is compact and metrizable, and $\pi_i^* : T_i^* \to \Delta(\Theta \times T_{-i}^*)$ is a homeomorphism for all i.

The infinite hierarchy of beliefs $q_i(t_i)$ associated to type t_i in the type space $\mathcal{T}$ is an element of T_i^*. Thus, q_i is a mapping from T_i to T_i^*. Let $Q_i(T) \subseteq T_i^*$ denote the range of q_i. Since Θ and T_i are compact metric spaces and π_i is continuous, the mapping q_i is continuous. Therefore, $Q_i(T)$ is compact. Moreover, q_i is a belief-preserving morphism in the sense that for any measurable $E \subseteq \Theta \times T_{-i}^*$

$$\pi_i^*(q_i(t_i))(E) = \pi_i(t_i)(\{(\theta, t_{-i}) : (\theta, q_{-i}(t_{-i})) \in E\}).$$

Finally, each q_i^z is also continuous, and hence each $Q_i^z(T)$ is compact.

We will call the type space $\mathcal{T}$ *simple* if distinct types have distinct first-order beliefs, that is, for each i, if t_i and t_i' are two distinct types in T_i, then $q_i^1(t_i) \neq q_i^1(t_i')$. In a simple type space, types of each individual can be described in terms of his first-order beliefs because T_i is homeomorphic to $Q_i^1(T)$. Indeed, since q_i^1 is continuous and injective, and T_i is compact, q_i^1 itself defines the homeomorphism from T_i to $Q_i^1(T)$.

Standard type spaces used in most applied work are simple. For instance, in a *payoff-type space*, (a) each individual knows her payoff type and has beliefs regarding the distribution of the payoff types of others as defined by some continuous function $b_i : \Theta_i \to \Delta\Theta_{-i}$, and (b) this environment is common knowledge. Thus, the payoff-type space is such that there is an injective continuous mapping $\hat{\theta}_i : T_i \to \Theta_i$ for all i, and for all types t_i and measurable subsets $E \subseteq \Theta_i \times \Theta_{-i} \times T_{-i}$,

$$\pi_i(t_i)(E) = b_i(\hat{\theta}_i(t_i))(\{\theta_{-i} : (\hat{\theta}_i(t_i), \theta_{-i}, \hat{\theta}_{-i}^{-1}(\theta_{-i})) \in E\}),$$

where $\hat{\theta}_{-i}^{-1}$ is the inverse of $\hat{\theta}_{-i}$. A payoff-type space is simple because distinct types have distinct payoff types. Often, a payoff-type space has the additional property that the individual beliefs b_i are derived from a common prior $b \in \Delta\Theta$. In this case, we refer to the resulting type space as a *common-prior payoff-type space.* In other instances, the common prior is itself assumed to be a product of independent distributions g_i over the payoff-type sets Θ_i, which we refer to as a *common-prior payoff-type space with independent types.* Sometimes values are assumed, in addition, to be *private*: $\Theta = \times_{i \in I} \Theta_i$ and only i's component of the state can impact his payoff. Another notable example of a simple type space is the *complete-information type space* $\mathcal{T}^{CI}$, where for all i, the set of types $T_i \subseteq \cup_{\theta \in \Theta} \{t_i^\theta\}$ and the belief $\pi_i(t_i^\theta)$ of each t_i^θ is such that $\pi_i(t_i^\theta)(\{\theta, t_{-i}^\theta\}) = 1$.

Our analysis focuses on problems where the planner's benchmark model is a simple type space (while robustness will be defined against any larger type space). A first reason for this is to make the analysis more relatable for people who are used to standard frameworks such as those discussed in the previous paragraph. A deeper reason is that, given our goal to accommodate bounded depths of reasoning, it seems unrealistic to rely on mechanisms asking participants to report high-order beliefs. Instead, we will restrict attention to simple mechanisms where the mechanism designer asks participants for their first-order beliefs (see next subsection). As a corollary, social choice functions implementable this way can only vary with first-order beliefs. It seems only more coherent then to assume that these first-order beliefs capture all the relevant information. That being said, we point out in the last section that our formal analysis extends well beyond the case of simple mechanisms and simple type spaces.

Planner's Model and 'Nearby' Type Profiles: The *planner's model* of the individuals' information is given by a simple type space $\hat{\mathcal{T}} = (\hat{T}_i, \hat{\pi}_i)_{i \in I}$. To formalize the notion of type profiles that are outside but 'nearby' the planner's model, the first step is to consider type spaces that include the planner's model. For any two type spaces $\mathcal{T}' = (T'_i, \pi'_i)_{i \in I}$ and $\mathcal{T} = (T_i, \pi_i)_{i \in I}$, we say $\mathcal{T}' \supseteq \mathcal{T}$ if for all i, $T'_i \supseteq T_i$, the set T_i is endowed with the relative topology induced by the topology on T'_i, and for all $t_i \in T_i$, $\pi'_i(t_i)(E) = \pi_i(t_i)(\Theta \times T_{-i} \cap E)$ for any measurable $E \subseteq \Theta \times T'_{-i}$. While $\hat{\mathcal{T}}$ is assumed to be simple, we do not impose any restriction on larger type spaces.

This will make continuous implementation more robust, and hence more demanding. Notice, however, that our results remain valid when imposing the larger type space to be simple as well.

The second step is to define what we mean by 'nearby'. There are two possible definitions. First, given $\mathcal{T}' \supseteq \mathcal{T}$, the sequence of type profiles $(t^n)_{n\geq 1}$ in T' *converges* to some $t \in T$ if $(t^n)_{n\geq 1}$ converges to t with respect to the topology on T'. We denote this by $t^n \to t$. Second, as in OT, the sequence of type profiles $(t^n)_{n\geq 1}$ in T' *converges^p* to some $t \in T$ if $(q(t^n))_{n\geq 1}$ converges to $q(t)$ with respect to the (product) topology on T^*, that is, for each $z \geq 1$ and i, the belief $q_i^z(t_i^n)$ converges to $q_i^z(t_i)$ in the weak* topology. We denote this by $t^n \xrightarrow{p} t$. Since each q_i^z is continuous, $t^n \to t$ implies $t^n \xrightarrow{p} t$. The converse is true whenever T' does not contain redundant types. We will use convergence^p to define 'nearby' profiles in the definition of continuous level-K implementation. This makes it easier to compare our results with those in OT. Moreover, using the weaker convergence^p notion imposes a more stringent requirement on the designer as she has to guarantee continuity of outcomes for a larger set of type-profile sequences converging to her model.

Recall that the universal type space T^* is metrizable. Let d_i^* be any metric consistent with the topology on T_i^*. Given any type space $\mathcal{T}'$, for each i define $d_i : T_i' \times T_i' \to \Re$ such that $d_i(t_i, t_i') = d_i^*(q_i(t_i), q_i(t_i'))$. Then d_i is a semimetric, and the topology induced by d_i on T_i' is called the *semimetric topology induced by* d_i. For any subset of individuals $I' \subseteq I$, we refer to the corresponding product topology on $\times_{i\in I'} T_i'$ as the *semimetric topology on* $\times_{i\in I'} T_i'$. It is easy to see that the notion of convergence^p is equivalent to convergence with respect to the semimetric topology on T'. In what follows, we will use the superscript "^{p}" whenever we want to make it clear that we are referring to these semimetric topologies.

Mechanisms: A *mechanism* is a measurable function $\mu : M \to \Delta X$, where $M = \times_{i\in I} M_i$ and each M_i is the set of messages m_i that player i can report. The mechanism μ and the type space $\mathcal{T}$ together define a Bayesian game, $\mathcal{G}(\mu, \mathcal{T})$, where individual i's *strategy* is a measurable function $\sigma_i : T_i \to \Delta M_i$. A strategy profile σ and type profile t induce a lottery $\mu(\sigma(t))$ over X.[7]

[7] Formally, for any Borel subset B of X, $\mu(\sigma(t))(B) = \int_M \mu(m)(B) d\sigma_1(t_1) \times \ldots \times \sigma_I(t_I)$.

Given a mechanism $\mu : M \to \Delta X$, type space $\mathcal{T}$, individual i, and strategy $\sigma_j : T_j \to M_j$ for $j \neq i$, the *best-response correspondence* $BR_i^{\sigma_{-i}} : T_i \to M_i$ is defined as

$$BR_i^{\sigma_{-i}}(t_i) = \arg\max_{m_i \in M_i} \int_{\Theta \times T_{-i}} U_i(\mu(m_i, \sigma_{-i}(t_{-i})), \theta) d\pi_i(t_i).$$

A *simple mechanism* is such that $M_i = \Delta\Theta$ for all i. Mechanisms so defined are "direct" in the sense that participants are asked to report their information,[8] thereby restricting attention to the most meaningful questions the social planner might ask. This assumption is made mostly to simplify the exposition; again, see our last section for a discussion.

Behavioral anchors – capturing individuals' gut reaction on how to play a game – play a key role in level-k models. That SIRBIC is sufficient for level-k implementation is straightforward with truthful anchors in direct mechanisms, but less so when considering other anchors such as uniform-random anchors over abstract messages (see dCSS). We find it then more effective to develop our arguments for continuous level-k implementation in the case of direct mechanisms and truthful anchors, and argue in the final section how our reasoning also applies to indirect mechanisms and other behavioral anchors. Another reason for our expositional choice is that level-k implementation with truthful anchors in direct mechanisms is equivalent to weak Bayesian implementation in strict equilibrium. Hence, the discrepancy arising between the two implementation concepts when adding the continuity requirement gets only more striking.

In addition to being direct, another feature of simple mechanisms is that participants report only beliefs about the state. This is without loss of generality when considering simple type spaces. As already mentioned earlier, asking higher-order beliefs seems less realistic when considering agents with bounded depth of reasoning, but the formal analysis does extend to nonsimple mechanisms and nonsimple type

[8] A *direct mechanism* is a mechanism such that $M_i = T_i$ for all i. In a simple type space, $Q_i(T)$ is homeomorphic to T_i, for all i. Hence, asking the players to report their first-order beliefs is equivalent to asking them to report their types. Nevertheless, a simple mechanism may accommodate more messages than truthful reports ($Q_i^1(T)$ may be a strict subset of $\Delta\Theta$ for some i). In that sense, simple mechanisms may not be direct mechanisms in the strict sense of that term. Alternatively, one could define simple mechanisms only over first-order beliefs that are compatible with the type space of interest. As this makes no difference to the analysis, we prefer a definition that is independent of any type space.

spaces (see concluding section).

Level-k Behavior: We fix a simple mechanism $\mu : (\Delta\Theta)^I \to \Delta X$ and a type space $\mathcal{T}$ (note that the type space $\mathcal{T}$ need not be simple, which allows us to define level-k behavior in general type spaces). For each $k \geq 1$, the level-k individual believes that all others are of level-$(k-1)$, and then best responds to their strategies. Behavioral hierarchies are then derived iteratively, starting with truth-telling as the behavioral anchor. Formally:

Definition 1. Individual i's strategy σ_i is *level*-1 *consistent* if it is a best response against others reporting the truth: for any type $t_i \in T_i$, messages in the support of $\sigma_i(t_i)$ maximize

$$\int_{\Theta\times T_{-i}} U_i\big(\mu(q_i^1, (q_j^1(t_j))_{j\neq i}), \theta\big) d\pi_i(t_i)$$

over $q_i^1 \in \Delta\Theta$. The set of all such strategies is denoted $S_i^1(\mu, \mathcal{T})$. By induction, for each $k \geq 2$, individual i's strategy σ_i is *level-k consistent* if it is a best response against a level-$(k-1)$ consistent strategy profile for the other individuals: for any type t_i, messages in the support of $\sigma_i(t_i)$ maximize $\int_{\Theta\times T_{-i}} U_i(\mu(q_i^1, \sigma_{-i}(t_{-i})), \theta) d\pi_i(t_i)$, for some $\sigma_{-i} \in S_{-i}^{k-1}(\mu, \mathcal{T})$. The set of all such strategies is denoted $S_i^k(\mu, \mathcal{T})$.

The index k is called an individual's *depth of reasoning*. Following the experimental literature, which suggests that individuals' depth of reasoning is bounded, usually by three or four levels (see the references in the Introduction), we will assume throughout the paper that each individual's depth of reasoning is bounded by some strictly positive integer K.[9] Although the experiments provide some evidence regarding the distribution of depths of reasoning in the population (viz., the proportion of individuals displaying a particular depth of reasoning decreases as the depth of reasoning increases), we do not have sufficient evidence to justify making any specific assumption regarding that distribution. We instead make the more robust assumption that the planner considers all combinations of depths of reasoning between 1 and the upper bound K as possible.

[9] Our results go through even if the upper bound on depths of reasoning is not the same across individuals. What is critical for the general necessary condition is that the upper bound for each individual is at least 2.

Continuous Level-k Implementation: Fix the planner's model $\hat{\mathcal{T}}$. She finds it desirable to implement a (measurable) *social choice function* (SCF) $f : \hat{T} \to \Delta X$. We say that an individual i is *irrelevant* for the SCF f if $f(t_i, t_{-i}) = f(t'_i, t_{-i})$, for all $t_i, t'_i \in \hat{T}_i$ and all $t_{-i} \in \hat{T}_{-i}$. Thus i's type matters under no circumstance when i is irrelevant. Individuals who are not irrelevant are called *relevant*. SCFs 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 the mechanism.

The planner wants to implement $f : \hat{T} \to \Delta X$ continuously, meaning that at each $t \in \hat{T}$, she wants the outcome $f(t)$ and an outcome close to $f(t)$ at type profiles near t. To achieve this, she constructs a simple mechanism μ, and assumes individuals play the resulting Bayesian game using some level-k consistent strategy with truthful anchors. We formalize this as follows.

Definition 2. The mechanism $\mu : (\Delta\Theta)^I \to \Delta X$ *continuously implements up to level-K* the SCF $f : \hat{T} \to \Delta X$ if the following two conditions are satisfied for all type spaces $\mathcal{T} \supseteq \hat{\mathcal{T}}$:

1. $S_i^k(\mu, \mathcal{T}) \neq \emptyset$, for all i and k such that $1 \leq k \leq K$.
2. Pick any sequences $(t^n)_{n\geq 1}$ in T such that $t_n \xrightarrow{p} t \in \hat{T}$. For any individual i, pick any k_i such that $1 \leq k_i \leq K$ and any strategy $\sigma_i \in S_i^{k_i}(\mu, \mathcal{T})$. Then the outcome sequence $\mu \circ \sigma(t^n)$ converges to $f(t)$.[10]

For a point of contrast, the planner could care only about achieving the SCF on $\hat{T}$ if she is confident in her benchmark model $\hat{\mathcal{T}}$. This corresponds to the notion of implementation studied in dCSS (in the special case of truthful anchors given a simple mechanism):

Definition 3. The mechanism $\mu : (\Delta\Theta)^I \to \Delta X$ *implements up to level-K* $f : \hat{T} \to \Delta X$ if the two conditions below are satisfied for the type space $\hat{\mathcal{T}}$:

1. $S_i^k(\mu, \hat{\mathcal{T}}) \neq \emptyset$, for all i and k such that $1 \leq k \leq K$.

[10] We do not require implementability for $k_i = 0$ because we view individuals as minimally rational in the sense of playing a best response to some belief. In that sense, there are no level-0 individuals, and behavioral anchors only capture individuals' beliefs regarding others' gut feelings towards the mechanism. In any case, results remain valid under truthful anchors when including $k_i = 0$ in the definition.

2. For any individual i, pick any k_i such that $1 \leq k_i \leq K$ and any strategy $\sigma_i \in S_i^{k_i}(\mu, \hat{\mathcal{T}})$. Then the outcome $\mu \circ \sigma(t) = f(t)$ for all $t \in \hat{T}$.

We shall refer to this notion as the merely exact implementation of f. Clearly, Definition 2 of continuous implementation implies Definition 3 of merely exact implementation (simply using constant sequences). Finally, we will assume throughout the paper that $K \geq 2$.

4 Characterization of Continuous Level-k Implementation

We begin the section by recalling the definition of SCFs satisfying SIRBIC. Pick any player i and a pair of types $t_i, t_i' \in \hat{T}_i$. 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 $t_{-i} \in \hat{T}_{-i}$. Otherwise, we say that f is *responsive* to t_i versus t_i'.

Definition 4. The SCF f is *strictly-if-responsive Bayesian incentive compatible (SIRBIC)* if

$$\int_{\Theta \times \hat{T}_{-i}} U_i(f(t), \theta) d\hat{\pi}_i(t_i) \geq \int_{\Theta \times \hat{T}_{-i}} U_i(f(t_i', t_{-i}), \theta) d\hat{\pi}_i(t_i), \tag{1}$$

for all $t_i, t_i' \in \hat{T}_i$ and $i \in I$, and the inequality holds strictly when f is responsive to t_i versus t_i'.

The above inequality means that each type of each individual wants to report his true type when everyone else reports their types truthfully in the direct mechanism associated with the SCF f. Additionally, the incentive to report truthfully must be strict whenever f is responsive to t_i versus t_i'. This additional requirement makes SIRBIC stronger than standard Bayesian incentive compatibility. Yet, as the incentive constraint must hold with equality whenever $t_i \sim_i^f t_i'$, SIRBIC is slightly weaker than strict Bayesian incentive compatibility. Recall the SCF f is *strictly Bayesian incentive compatible* if the inequality (1) is strict for all $t_i' \neq t_i$ and $i \in I$.

The main results in dCSS imply that f is implementable up to level-K using a simple mechanism if and only if f satisfies SIRBIC. Note that the restriction to simple mechanisms is not required there. But, under truthful anchors, sufficiency was proved by using the SCF as the direct mechanism. With a simple type space, we can easily adjust the construction to obtain a simple mechanism implementing f: for each $q_i^1 \in \Delta\Theta$, let $\hat{q}_i^1 = q_i^1$ if $q_i^1 \in Q_i^1(\hat{T})$ and $\hat{q}_i^1 = q_i^{1*}$ for some $q_i^{1*} \in Q_i^1(\hat{T})$ if $q_i^1 \notin Q_i^1(\hat{T})$; then define $\mu(q_1^1, \ldots, q_I^1) = f\big((\tau_i(\hat{q}_i^1))_{i\in I}\big)$ where $\tau_i(\hat{q}_i^1)$ is the unique type in $\hat{T}_i$ associated to the first-order belief $\hat{q}_i^1$. With truth-telling at level-0, SIRBIC provides the incentives to individuals at level-1 and higher to report either their true first-order belief $q_i^1(t_i)$ or a q_i^1 that is mapped back into the first-order belief of a type $t_i' \sim^f t_i$, outcome-equivalent to their true type.

Our main finding in the current paper is that the characterization result in dCSS extends to continuous level-k implementation, with the only addition that f must be continuous:

Theorem 1. *The SCF $f : \hat{T} \to \Delta X$ is continuously implementable up to level-K if and only if f is continuous and satisfies SIRBIC.*[11]

Proof. (Necessity) As noted at the end of the previous section, continuous implementation up to level-K implies merely exact implementation up to level-K. By Theorem 1 in dCSS, f satisfies SIRBIC.[12] To argue that f is continuous, consider any sequence $(t^n)_{n\geq 1}$ in $\hat{T}$ converging to some $t \in \hat{T}$. Then $t^n \xrightarrow{p} t$. We show that $f(t^n)$ converges to $f(t)$ using the mechanism μ that continuously implements f. For each individual i, pick a level-1 consistent strategy $\sigma_i^1 \in S_i^1(\mu, \hat{\mathcal{T}})$. Then the fact that μ continuously implements up to level-K the SCF f implies that $\mu \circ \sigma^1(t^n) = f(t^n)$ for all n and the sequence of outcomes $\mu \circ \sigma^1(t^n)$ converges to $f(t)$. Thus, $f(t^n)$ converges to $f(t)$.

(Sufficiency) This proof is quite technical. Here we provide only an outline of its main steps, relegating some important auxiliary lemmatta to the Appendix, for ease of exposition.

Since $\hat{\mathcal{T}}$ is simple, $q_i^1 : \hat{T}_i \to Q_i^1(\hat{T})$ is a homeomorphism. Let τ_i be the inverse of q_i^1. For any $q^1 \in \times_{i\in I} Q_i^1(\hat{T})$, let $\tau(q^1) = (\tau_1(q_1^1), \ldots, \tau_I(q_I^1))$. Define $\hat{\mu} : \times_{i\in I} Q_i^1(\hat{T}) \to \Delta X$

[11] A very early and incomplete precursor of this result first appeared in our unpublished working paper, de Clippel *et al.* (2014).

[12] dCSS proves this necessary condition while assuming that the planner's model is given by a common-prior payoff-type space. Although we are allowing for more general type spaces here, essentially the same arguments apply as well.

as

$$\hat{\mu}(q^1) = f(\tau(q^1)), \forall q^1 \in \times_{i\in I} Q_i^1(\hat{T}).$$

If $Q_i^1(\hat{T}) = \Delta\Theta$ for each i, this defines a simple mechanism that continuously implements f up to level-K. Indeed, just like for merely exact implementation discussed above, SIRBIC implies that types in the planner's model, whatever their depths of reasoning, report either their true first-order belief $q_i^1(t_i)$ or a q_i^1 that is mapped back into the first-order belief of a type $t_i' \sim^f t_i$, outcome-equivalent to their true type. Then the second condition in Definition 2 obtains from the facts that f is continuous and the correspondence Σ_i^k (Definition 5), which contains the level-k correspondence, is upper hemicontinuous^p, as discussed below .

But things get more complicated when $Q_i^1(\hat{T})$ is a *strict* subset of $\Delta\Theta$ for some i (as, e.g., when $\hat{T}$ describes any payoff-type space, or a complete-information environment). Now the function $\hat{\mu}$ must be extended into a simple mechanism μ defined over the larger set $(\Delta\Theta)^I$. While the continuity of f guarantees that $\hat{\mu}$ is continuous, and hence that one can find a continuous extension μ, the fact that the level-k correspondence

$$t_i \to \bigcup_{\sigma_i \in S_i^k(\mu,\mathcal{T})} \text{support}[\sigma_i(t_i)] \tag{2}$$

need not be upper hemicontinuous (see Section 2) means that not all continuous extensions of $\hat{\mu}$ will continuously implement f up to level-K.[13] The difficulty is to establish the second condition in Definition 2 when elements of $Q^1(\hat{T})$ are approached by sequences of beliefs outside this set, which of course no longer follows from the continuity of f, but instead requires using a thoughtful extension μ.

Example 1 highlights the main reason why the level-k correspondence could fail to be upper hemicontinuous: an individual may believe under the enlarged type space that others' behavior correlates with the state, but not in the planner's model. For this reason, we construct for each individual i and depth of reasoning $k \geq 1$ the correspondence Σ_i^k that associates to each of his types, individual i's best-response messages against all conjectures in $\Delta(\Theta \times T_{-i} \times (\Delta\Theta)^{I-1})$ 'consistent' with the behavior of level-$(k-1)$ opponents (with truth-telling at level-0). Thus, unlike S_i^k,

[13] A lack of upper hemicontinuity also implies a lack of upper hemicontinuity^p given that $t_i^n \to t_i$ implies $t_i^n \overset{p}{\to} t_i$.

the definition of Σ_i^k allows individuals to perceive θ, t_{-i}, and other's messages as correlated. Formally:

Definition 5. Given the mechanism $\mu : (\Delta\Theta)^I \to \Delta X$, and letting α_j denote j's truth-telling strategy, define $\Sigma_i^1(t_i|\mu, \mathcal{T}) = BR_i^{\alpha_{-i}}(t_i)$, for all $t_i \in T_i$, and for each $k \geq 2$, inductively define $\Sigma_i^k(t_i|\mu, \mathcal{T})$ as the union of the sets

$$\arg\max_{m_i \in \Delta\Theta} \int_{\Theta \times T_{-i} \times (\Delta\Theta)^{I-1}} U_i(\mu(m_i, m_{-i}), \theta) d\gamma$$

over all conjectures $\gamma \in \Delta(\Theta \times T_{-i} \times (\Delta\Theta)^{I-1})$ such that *(a)* the distribution $\pi_i(t_i)$ coincides with the marginal distribution of γ on $\Theta \times T_{-i}$, and *(b)* the marginal distribution of γ on $T_{-i} \times (\Delta\Theta)^{I-1}$ supports a subset of $\times_{j \neq i} Gr(\Sigma_j^{k-1}(\cdot|\mu, \mathcal{T}))$, where $Gr(\Sigma_j^{k-1}(\cdot|\mu, \mathcal{T}))$ is the graph of $\Sigma_j^k(\cdot|\mu, \mathcal{T})$.

Then, in a nutshell, the general proof of sufficiency proceeds by identifying an adequate continuous extension μ of $\hat{\mu}$, establishing the upper hemicontinuity^p of $\Sigma_i^k(\cdot|\mu, \mathcal{T})$, which contains the (not necessarily upper hemicontinuous) correspondence defined in (2), and combining these different pieces in order to establish the second condition in Definition 2.

Specifically, we define the extension μ by applying Lemma 1 from the Appendix. This lemma is a variant of Dugundji (1951) and shows how to extend a continuous mechanism defined on a closed subset of a compact metric message space to the whole message space in such a manner that the resulting extended mechanism is continuous. Applied to the current context, Lemma 1 shows that we can continuously extend $\hat{\mu}$ after "translating" each message in $\Delta\Theta \setminus Q_i^1(\hat{T})$ into a finite probability distribution over messages in $Q_i^1(\hat{T})$, while keeping messages in $Q_i^1(\hat{T})$ unchanged.[14]

[14] At first blush, an obvious choice would be to use a single-valued selection of the projection operator for this translation. Unfortunately, one cannot guarantee the continuity of the resulting extended mechanism without additional conditions on $Q_i^1(\hat{T})$. Continuity does obtain, however, if one uses our more elaborate construction. Although one could take a host of alternative approaches to obtain the continuous extension (e.g., apply Dugundji's result as is, if we overlook the product structure, or apply his result component by component), it is more appropriate to provide a new construction. Indeed, with respect to Dugundji (1951), our version differs from the former two approaches in that the probability of picking a message profile q^1 is the product of probabilities with each factor depending *only* on i's reported belief. This kind of product/separability property is needed, e.g., in Lemma 6 of the proof.

The next step is to establish that level-k behavior in μ is in some sense "continuous" in types. Although an arbitrary level-k consistent strategy is not necessarily continuous and the level-k correspondence defined in (2) is not necessarily upper hemicontinuous (as discussed above), the latter is a selection of $\Sigma_i^k(\cdot|\mu, \mathcal{T})$, which is upper hemicontinuous (Lemmata 3 and 4). To conclude these more technical steps of the proof, Lemma 5 establishes that $\Sigma_i^k(\cdot|\mu, \mathcal{T})$ depends only on types' belief hierarchies of types, which implies that the correspondence is in fact upper hemicontinuous^p.

The final important step in the proof is to show that for all type spaces $\mathcal{T} \supseteq \hat{T}$, individuals i, depths of reasoning $k \geq 1$, and types t_i who belong to the planner's model, the constructed mechanism μ is such that messages in $\Sigma_i^k(t_i|\mu, \mathcal{T})$ translate into messages in $\{q_i^1(t_i') : t_i' \sim_i^f t_i\}$ (Lemma 6). By itself, the SIRBIC property of f only guarantees that when individuals play the underlying mechanism $\hat{\mu}$ on the planner's model and the behavioral anchors are truthful, then $\Sigma_i^k(t_i|\hat{\mu}, \hat{\mathcal{T}}) = \{q_i^1(t_i') : t_i' \sim_i^f t_i\}$ for all $k \geq 1$ and $t_i \in \hat{T}_i$. The specific way we extend $\hat{\mu}$ to μ then allows us to preserve the incentives of the types in the planner's model to ensure that they only want to report those messages in $\Delta\Theta$ that translate as equivalent to truth-telling. This property, paired with the upper hemicontinuity^p of $\Sigma_i^k(\cdot|\mu, \mathcal{T})$, then implies that the extended mechanism μ continuously implements f up to level-K. □

5 Equilibrium versus Level-k in Continuous Implementation

One key task in our effort is to evaluate how different the conditions for continuous level-k implementation are from those that yield continuous implementation in strict Bayesian equilibrium (as defined in OT), the latter well-known to be very restrictive. In the current section, we disentangle the properties implied by either type of implementation. This exercise is useful to find that neither continuity *per se* nor SIRBIC (which has to be satisfied for types in the planner's model) are responsible for the very restrictive results in OT. Rather, it is the insistence on the use of Bayesian equilibrium, its corresponding incentive constraints in the larger model, that is the culprit.

Given a (not necessarily simple) mechanism μ, the strategy profile σ is a *strict Bayes Nash equilibrium* (strict BNE) in the game $\mathcal{G}(\mu, \hat{\mathcal{T}})$ if $BR_i^{\sigma_{-i}}(t_i) = \{\sigma_i(t_i)\}$, for all $i \in I$ and $t_i \in \hat{T}_i$. The mechanism μ *strictly equilibrium-implements* $f : \hat{T} \to \Delta X$ if $\mathcal{G}(\mu, \hat{\mathcal{T}})$ admits a strict BNE σ such that $f(\hat{t}) = \mu \circ \sigma(\hat{t})$, for all $\hat{t} \in \hat{T}$.

Observe that f is strictly equilibrium-implementable if and only if it satisfies SIRBIC. Sufficiency is obvious (simply apply f as the outcome function, after keeping only one representative of each $\sim_i^f$-equivalence class). As for necessity, consider any t_i and t_i' in $\hat{T}_i$ such that $t_i \not\sim_i^f t_i'$. Then it must be that $\sigma_i(t_i) \neq \sigma_i(t_i')$, and since σ is a strict BNE in $\mathcal{G}(\mu, \hat{\mathcal{T}})$, it must be that

$$\int_{\Theta \times \hat{T}_{-i}} U_i(\mu(\sigma(t)), \theta) d\hat{\pi}_i(t_i) > \int_{\Theta \times \hat{T}_{-i}} U_i(\mu(\sigma_i(t_i'), \sigma_{-i}(t_{-i})), \theta) d\hat{\pi}_i(t_i),$$

and hence that

$$\int_{\Theta \times \hat{T}_{-i}} U_i(f(t), \theta) d\hat{\pi}_i(t_i) > \int_{\Theta \times \hat{T}_{-i}} U_i(f(t_i', t_{-i}), \theta) d\hat{\pi}_i(t_i).$$

Thus, strict BNE implementation and implementation up to level-K are both equivalent to SIRBIC. This perhaps surprising coincidence is explored in greater depth in dCSS. The present paper raises a new puzzle: continuous implementation obtains almost for free under level-k behavior (Theorem 1 above), but can be very restrictive under strict BNE (see OT). This section attempts to shed light on this surprising difference.

Next, we provide the definition of strict BNE for extended types spaces, towards formulating continuous implementation, as follows:

Given a (not necessarily simple) mechanism μ and type space $\mathcal{T} \supseteq \hat{T}$, the strategy profile σ is a *Bayes Nash equilibrium* (BNE) in the game $\mathcal{G}(\mu, \mathcal{T})$ if for all $i \in I$ and $t_i \in T_i$, any message m_i in the support of $\sigma_i(t_i)$ belongs to $BR_i^{\sigma_{-i}}(t_i)$. Let $\sigma|_{\hat{T}}$ denote the restriction of σ to the domain $\hat{T}$. The BNE σ in $\mathcal{G}(\mu, \mathcal{T})$ *continuously strictly implements* $f : \hat{T} \to \Delta X$ if (i) $\sigma|_{\hat{T}}$ is a strict BNE in $\mathcal{G}(\mu, \hat{T})$ and (ii) for any $t \in \hat{T}$ and any sequence $(t^n)_{n \geq 1}$ in T such that $t^n \xrightarrow{p} t$, we have $\mu \circ \sigma(t^n) \to f(t)$. Following OT, the SCF $f : \hat{T} \to \Delta X$ is *continuously strictly equilibrium-implementable* if there exists a mechanism μ such that for all type spaces $\mathcal{T} \supseteq \hat{\mathcal{T}}$, there exists an equilibrium σ in $\mathcal{G}(\mu, \mathcal{T})$ that strictly continuously equilibrium-implements f. We now essentially apply the revelation principle to obtain a necessary condition for continuous strict equilibrium-implementation, and highlight a key difference with continuous level-k implementation:

Theorem 2. *If $f : \hat{T} \to \Delta X$ is continuously strictly equilibrium-implementable, then for any type space $\mathcal{T} \supseteq \hat{\mathcal{T}}$, there exists an SCF $g : T \to \Delta X$ such that*

(i) g is a continuous^p extension of f:

If $t^n \xrightarrow{p} t \in \hat{T}$ for some sequence $(t^n)_{n\geq 1}$ in T, then $g(t^n) \to f(t)$.

(ii) g satisfies SIRBIC with respect to types in $\hat{T}$:

$$\int_{\Theta\times\hat{T}_{-i}} U_i(g(t),\theta)d\hat{\pi}_i(t_i) \geq \int_{\Theta\times\hat{T}_{-i}} U_i(g(t'_i,t_{-i}),\theta)d\hat{\pi}_i(t_i), \tag{3}$$

for all $t_i \in \hat{T}_i$, $t'_i \in T_i$, and $i \in I$, with the inequality holding strictly when $t_i \not\sim_i^g t'_i$.

(iii) g is Bayesian incentive compatible:

$$\int_{\Theta\times T_{-i}} U_i(g(t),\theta)d\pi_i(t_i) \geq \int_{\Theta\times T_{-i}} U_i(g(t'_i,t_{-i}),\theta)d\pi_i(t_i), \tag{4}$$

for all $t_i, t'_i \in T_i$ and $i \in I$.

In contrast, f is continuously implementable up to level-K if and only if for any type space $\mathcal{T} \supseteq \hat{\mathcal{T}}$, there exists $g : \mathcal{T} \to \Delta X$ such that (i) and (ii) hold.

Before proving the theorem, let us briefly describe what the properties mean. Property (i) is self-explanatory: $g(t^n)$ is as close as desired to $f(t)$ if n is large enough. It implies in particular that f and g coincide on $\hat{T}$ (simply consider a constant sequence in $\hat{T}$ to check this). Inequality (4) says that individual i of type $t_i \in T_i$ prefers to report his type truthfully in the direct mechanism associated with the SCF g whenever he believes that the other individuals also report their respective types truthfully. Inequality (3) requires that as well, this time for types $t_i \in \hat{T}_i$, but asks this preference to be strict whenever g is responsive to t_i versus any $t'_i \in T_i$. Notice that this is more demanding than SIRBIC of g restricted to $\hat{T}$ (which equals f under (i)), as t_i' need not belong to $\hat{T}_i$. Finally, notice the order of quantifiers: g can vary with $\mathcal{T}$.

Proof. (*Continuous Strict Equilibrium-Implementation*) Let μ be any mechanism. Consider a type space $\mathcal{T} \supseteq \hat{\mathcal{T}}$, and let σ be an equilibrium in $\mathcal{G}(\mu, \mathcal{T})$ that strictly continuously implements f. Then $g : \mathcal{T} \to \Delta X$, defined by $g(t) = \mu \circ \sigma(t)$ for all $t \in \mathcal{T}$, is a continuous^p extension of f: for any $t \in \hat{T}$ and any sequence $(t^n)_{n\geq 1}$ in $\mathcal{T}$ such that $t^n \xrightarrow{p} t$, we have $g(t^n) = \mu \circ \sigma(t^n) \to f(t)$. By the standard revelation principle, g is Bayesian incentive compatible. Finally, the same argument, presented at the beginning of this section to establish that SIRBIC is a necessary condition for strict equilibrium-implementation, implies that g satisfies SIRBIC with respect to types in $\hat{T}$.

(*Continuous Implementation Up To level-K*) Starting with sufficiency, using g associated to $\mathcal{T} = \hat{\mathcal{T}}$, we get from (i) that $f = g$ is continuous and from (ii) that f satisfies SIRBIC. By Theorem 1, f is continuously implementable up to level-K. As for necessity, suppose f is continuously implementable up to level-K. By Theorem 1, it is continuous and satisfies SIRBIC. Hence, it can be continuously implemented up to level-K using the continuous extension μ of $\hat{\mu}$ presented in the proof of Theorem 1. Define $g(t) = \mu\big(q^1(t)\big)$ for all $t \in \mathcal{T} \supseteq \hat{\mathcal{T}}$. Then g is a continuous^p extension of f: for any $t \in \hat{T}$ and any sequence $(t^n)_{n\geq 1}$ in T such that $t^n \xrightarrow{p} t$, we have $q^1(t^n) \to q^1(t)$, and hence, by the continuity of μ, we have $g(t^n) = \mu\big(q^1(t^n)\big) \to \mu\big(q^1(t)\big)$. But $\mu\big(q^1(t)\big) = \hat{\mu}\big(q^1(t)\big) = f(t)$, by the construction of μ. Also, g satisfies SIRBIC with respect to types in $\hat{T}$ because the construction of μ is such that each $t_i \in \hat{T}_i$ wants to report only those messages in $\Delta\Theta$ that translate as equivalent to $q_i^1(t_i)$ when everyone else is reporting their first-order beliefs truthfully (Lemma 6). □

Conditions in Theorem 2 are phrased in terms of SCFs, with no direct reference to strict Bayes equilibria or level-k behavior in mechanisms. This facilitates comparisons between the two notions of continuous implementation. While we do not know whether (i)-(iii) in Theorem 2 characterizes strict continuous equilibrium-implementation, we can show that (iii) is indeed a key distinction when comparing the equilibrium and level-k approaches to continuous implementation. We start by discussing an example, and then prove that (i)-(iii) imply strict interim rationalizable monotonicity when considering a discrete type space for the planner, as defined in OT. This condition, which boils down to a strenghthening of Maskin monotonicity for complete-information environments, is very restrictive. We remark that we find no such connection to monotonicity conditions in the continuous level-k analysis, from which it is apparent that item (iii) in Theorem 2 is the culprit for the very restrictive results in OT. We turn to the example next.

Example 2. Consider a simple bilateral trade problem where the good to be traded can be of low (θ_L) or high (θ_H) quality. The buyer's reservation price is \$50 (*resp.* \$60) if the good is of low (*resp.* high) quality, while the seller's reservation price is fixed at \$0 irrespective of the good's quality. An alternative specifies $p \in \{0, 1\}$, where $p = 1$ means that the good is traded whereas $p = 0$ means that it is not traded, and a payment $z \in [-z^*, z^*]$ from the buyer to the seller, where z^* is sufficiently large.

The designer's model $\hat{T}$ is one of complete information, which is a simple type space. That is, $\hat{T}_i = \{t_i^L, t_i^H\}$ for all $i = b, s$, where b denotes the buyer and s the seller, and the beliefs are

$$\hat{\pi}_i(t_i^L)(\theta_L, t_{-i}^L) = 1 \text{ and } \hat{\pi}_i(t_i^H)(\theta_H, t_{-i}^H) = 1, \forall i.$$

Consider the following SCF f on $\hat{T}$:

	t_s^L	t_s^H
t_b^L	Trade at price \$25	No trade and payment
t_b^H	No trade and payment	Trade at price \$30

This SCF is strictly Bayesian incentive compatible, and hence strictly equilibrium-implementable using the associated direct mechanism, satisfies SIRBIC, and is trivially continuous. Thus, it is continuously implementable up to level-K. But it fails to be continuously strictly equilibrium-implementable because it is impossible to extend the SCF to all larger type spaces, while simultaneously satisfying (i)-(iii) in Theorem 2, as discussed below.

A Simple Extension Satisfying (i) and (ii), but not (iii)

Of course, we know that an extension of f satisfying (i) and (ii) exists, by Theorem 1, and one could follow its sufficiency proof to find one. It will follow from our next point that this extension violates (iii). But it may be insightful to see a simpler, direct extension when it comes to this particular example. For each type t (in any extension $\mathcal{T}$ of $\hat{\mathcal{T}}$), define $g(t)$ as a lottery where the good is traded for \$25 with probability $q_b^1(t_b)(\theta_L)q_s^1(t_s)(\theta_L)$, the good is traded for \$30 with probability $(1-q_b^1(t_b)(\theta_L))(1-q_s^1(t_s)(\theta_L))$, and there is no trade and payment with the remaining probability. Notice that g is a continuous^p extension of f, because $t^n \xrightarrow{p} t$ implies that $q_i^1(t_i^n)(\theta_L)$ converges to $q_i^1(t_i)(\theta_L)$ for all i. It is also SIRBIC with respect to types in $\hat{T}$. For example, if the seller reports his type truthfully in the associated direct mechanism, then type t_b^L of the buyer expects to trade at price \$25 by reporting t_b^L or any other t_b such that $q_b^1(t_b)(\theta_L) = 1$. Note that $t_b \sim_b^g t_b^L$ if and only if $q_b^1(t_b)(\theta_L) = 1$. If type t_b^L reports t_b such that $q_b^1(t_b)(\theta_L) < 1$, then he expects either to trade at price \$25 with probability $q_b^1(t_b)(\theta_L)$ or not trade with probability $(1 - q_b^1(t_b)(\theta_L))$, which is strictly worse than the outcome under truth-telling.

The above extension, however, is not Bayesian incentive compatible on all larger type spaces (item (iii) in Theorem 2). To see this, consider, for example, the type space $\mathcal{T}$ such that $T_s = \{t_s^L, t_s^H, t_s\}$ and $T_b = \hat{T}_b$, where the beliefs of t_i^L and t_i^H are the same as before, while the belief of t_s is such that $\pi_s(t_s)(\theta_L, t_b^L) = \pi_s(t_s)(\theta_L, t_b^H) = 0.5$. If the buyer reports truthfully in the direct mechanism associated with g, then a truthful type t_s expects trade at \$25 or no trade with equal probability, which is worse than misreporting his type as t_s^H (in which case there is an equal chance of trade at \$30, or no trade).

No Extension Satisfying (i)-(iii)

But of course, g is but one example of extension. Could there be ways of extending f to all larger type spaces, while simultaneously satisfying the three conditions in Theorem 2? We now prove this is impossible, thereby illustrating how insisting on (iii) can make a big difference. Consider the type space $\mathcal{T}$ such that $T_i = \hat{T}_i \bigcup_{n\geq 2}\{t_i^n\}$ for all i, where the beliefs of t_i^L and t_i^H are the same as before, while other types' beliefs are: $\pi_i(t_i^2)(\theta_L, t_j^L) = \pi_i(t_i^2)(\theta_H, t_j^L) = 0.5$, and for all $n > 2$, $\pi_i(t_i^n)(\theta_L, t_j^{n-1}) = \frac{1}{n}$ and $\pi_i(t_i^n)(\theta_H, t_j^{n-1}) = 1 - \frac{1}{n}$. Notice that $\mathcal{T}$ is a simple type space. We impose the following metric on this type space: $d_i(t_i, t'_i) = |q^1_i(t_i)(\theta_H) - q^1_i(t'_i)(\theta_H)|$ for all $t_i, t'_i \in T_i$. Under the topology induced by this metric, $t^n_i \to t^H_i$ and π_i is continuous.

Suppose there exists an extension g that satisfies (ii) and (iii), i.e., it is SIRBIC with respect to types in $\hat{T}$ and Bayesian incentive compatible. The SCF g specifies the probability of trade $p(t_b, t_s)$ and the expected payment from the buyer to the seller $z(t_b, t_s)$ for all $(t_b, t_s) \in T$. We use induction to show that $t^n_i \sim^g_i t^L_i$ for all i and $n \geq 2$.

First consider $n = 2$. Types t^2_s and t^L_s of the seller believe that the buyer's type is t^L_b. Hence, (iii), i.e., Bayesian incentive compatibility implies that type t^L_s weakly prefers $g(t^L_b, t^L_s)$ to $g(t^L_b, t^2_s)$, while type t^2_s weakly prefers $g(t^L_b, t^2_s)$ to $g(t^L_b, t^L_s)$. But all types of the seller have the same preferences over alternatives. Hence, type t^L_s must be indifferent between $g(t^L_b, t^L_s)$ and $g(t^L_b, t^2_s)$, and (ii), i.e., SIRBIC with respect to types in $\hat{T}$ implies that $t^2_s \sim^g_s t^L_s$. As for the buyer, types t^2_b and t^L_b believe that the seller's type is t^L_s. Hence, (iii) implies that type t^L_b weakly prefers $g(t^L_b, t^L_s)$ to $g(t^2_b, t^L_s)$ while type t^2_b weakly prefers $g(t^2_b, t^L_s)$ to $g(t^L_b, t^L_s)$. The former implies that $(1 - p(t^2_b, t^L_s))50 \geq 25 - z(t^2_b, t^L_s)$, while the latter implies that $25 - z(t^2_b, t^L_s) \geq (1 - p(t^2_b, t^L_s))55 \geq (1 - p(t^2_b, t^L_s))50$. Hence, we must have $z(t^2_b, t^L_s) = 25$ and $p(t^2_b, t^L_s) = 1$, which makes t^L_b indifferent between $g(t^L_b, t^L_s)$ and $g(t^2_b, t^L_s)$. Then (ii) implies that $t^2_b \sim^g_b t^L_b$.

Next, suppose the statement is true for $n - 1$, where $n \geq 3$. We now argue that the statement must also hold for n. For the seller, type t^n_s of the seller believes that the buyer's type is t^{n-1}_b. Hence, (iii) implies that type t^n_s weakly prefers $g(t^{n-1}_b, t^n_s)$ to $g(t^{n-1}_b, t^{n-1}_s)$. Since $t^{n-1}_i \sim^g_i t^L_i$ for all i, we have $g(t^{n-1}_b, t^n_s) = g(t^L_b, t^n_s)$ and $g(t^{n-1}_b, t^{n-1}_s) = g(t^L_b, t^L_s)$. As all types of the seller have the same preferences over alternatives, type t^L_s must weakly prefer $g(t^L_b, t^n_s)$ to $g(t^L_b, t^L_s)$. But (iii) requires that type t^L_s weakly prefer $g(t^L_b, t^L_s)$ to $g(t^L_b, t^n_s)$. Hence, type t^L_s must in fact be indifferent between $g(t^L_b, t^L_s)$ and $g(t^L_b, t^n_s)$. Then (ii) implies that $t^n_s \sim^g_s t^L_s$. As for the buyer, type t^n_b believes that the seller's type is t^{n-1}_s. Hence, (iii) implies that type t^n_b weakly prefers $g(t^n_b, t^{n-1}_s)$ to $g(t^{n-1}_b, t^{n-1}_s)$. Since $t^{n-1}_i \sim^g_i t^L_i$ for all i, we have $g(t^n_b, t^{n-1}_s) = g(t^n_b, t^L_s)$ and $g(t^{n-1}_b, t^{n-1}_s) = g(t^L_b, t^L_s)$. So t^n_b weakly prefers $g(t^n_b, t^L_s)$ to $g(t^L_b, t^L_s)$, which implies that $25 - z(t^n_b, t^L_s) \geq (1 - p(t^n_b, t^L_s))\left(60 - \frac{10}{n}\right) \geq (1 - p(t^n_b, t^L_s))50$. Hence, type t^L_b must weakly prefer $g(t^n_b, t^L_s)$ to $g(t^L_b, t^L_s)$. But (iii) requires that type t^L_b weakly prefers

$g(t_b^L, t_s^L)$ to $g(t_b^n, t_s^L)$. Hence, type t_b^L must in fact be indifferent between $g(t_b^L, t_s^L)$ and $g(t_b^n, t_s^L)$. Then (ii) implies that $t_b^n \sim_b^g t_s^L$.

The fact that $t_i^n \sim_i^g t_i^L$ for all i and n implies that $g(t_b^n, t_s^n) = g(t_b^L, t_s^L)$ for all n. Thus, $t^n \xrightarrow{p} t^H$ (because by construction $t^n \to t^H$) but $g(t^n)$ does not converge to $g(t^H)$. In other words, g is not continuous^p at t^H, and hence fails (i).

We see in the example above how the combination of (i)-(iii) in Theorem 2 is indeed demanding, and more so than having just (i) and (ii). But could it be that this is a somewhat unique example, and that adding (iii) does not make much difference in general? The next result shows that condition (iii) does play a key role in making continuous strict equilibrium-implementation much more demanding than its level-k counterpart. A key result in OT (see their Theorem 3) is that strict interim rationalizable monotonicity (strict IRM) is necessary for continuous strict equilibrium-implementation. Using the same assumptions as in OT, in our last result, we show next that strict IRM is also a necessary condition for satisfying (i) to (iii) in Theorem 2. To introduce the result, we present the necessary definitions:

Matching the setting in OT, assume that the state space Θ and the allowed type spaces $\mathcal{T}$ are countable.

A *deception* is a collection of correspondences $\beta = (\beta_i)_{i \in I}$, where $\beta_i : \hat{T}_i \to \hat{T}_i$. We let $\hat{T}_i^\beta$ denote the range of β_i, i.e., $\hat{T}_i^\beta = \cup_{t_i \in \hat{T}_i} \beta_i(t_i)$. A deception is *acceptable* if $f(\tilde{t}) = f(t)$ for all $t \in \hat{T}$ and $\tilde{t} \in \beta(t)$; otherwise, the deception is *unacceptable.*

Definition 6. The SCF $f : \hat{T} \to \Delta X$ satisfies *interim rationalizable monotonicity (IRM)* if for every unacceptable deception there exist $i \in I$, $t_i \in \hat{T}_i$, and $\tilde{t}_i \in \beta_i(t_i)$ such that for every $\gamma \in \Delta(\Theta \times \hat{T}_{-i} \times \hat{T}_{-i}^\beta)$ such that the marginal distribution of γ on $\Theta \times \hat{T}_{-i}$ coincides with $\hat{\pi}_i(t_i)$ and $\gamma(\theta, t_{-i}, \tilde{t}_{-i}) > 0 \Longrightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$, there exists $\ell : \hat{T}_{-i} \to \Delta X$ such that

$$\sum_{\Theta \times \hat{T}_{-i}^\beta} U_i(\ell(t_{-i}), \theta) \operatorname{marg}_{\Theta \times \hat{T}_{-i}^\beta} \gamma(\theta, t_{-i}) > \sum_{\Theta \times \hat{T}_{-i}^\beta} U_i(f(\tilde{t}_i, t_{-i}), \theta) \operatorname{marg}_{\Theta \times \hat{T}_{-i}^\beta} \gamma(\theta, t_{-i}),$$

and

$$\sum_{\Theta \times \hat{T}_{-i}} U_i(f(t_i', t_{-i}), \theta) \hat{\pi}_i(t_i')(\theta, t_{-i}) \geq \sum_{\Theta \times \hat{T}_{-i}} U_i(\ell(t_{-i}), \theta) \hat{\pi}_i(t_i')(\theta, t_{-i}),$$

for all $t_i' \in \hat{T}_i$.

If the last inequality is strict whenever $t_i' = \tilde{t}_i$, then the SCF f satisfies *strict interim rationalizable monotonicity (strict IRM).*

After this definition, we are ready to state our last result, whose proof can also be found in the Appendix:

Theorem 3. *Suppose the SCF $f : \hat{T} \to \Delta X$ can be extended to every type space $\mathcal{T} \supseteq \hat{\mathcal{T}}$ such that the corresponding extension g is a continuous^p extension of f, satisfies SIRBIC with respect to types in $\hat{T}$, and is Bayesian incentive compatible. Then f satisfies strict interim rationalizable monotonicity.*

6 Discussion: Generalizations and Applications

By imposing a bound on the agents' depth of reasoning, which we assume starts with truthful behavioral anchors, we have presented results to show the permissiveness of mechanism design. In spite of requiring full implementation, only continuity and a simple strenghthening of incentive compatibility (SIRBIC) are in many settings the limitations to locally robust implementation with bounded depth of reasoning.

The result is far from obvious. In particular, one needs to overcome an important difficulty, i.e., that the iterated best-response correspondence associated with level-k reasoning may fail to be upper hemicontinuous, as detailed in Section 2. An important subtlety created by this fact for level-k mechanism design –in comparison with approaches that insist on common belief of rationality– is the following. Di Tillio (2011) shows that if a (finite) mechanism fully implements an SCF in rationalizable strategies, then the same mechanism continuously and fully implements the SCF in rationalizable strategies. The argument follows easily from the upper hemicontinuity of the rationalizability correspondence in that case (Dekel et al., 2007). Significantly, the failure of upper hemicontinuity of the level-k correspondence implies that the same result is not necessarily true for level-k implementation, that is, if a mechanism level-k implements an SCF, then the same mechanism need not continuously level-k implement the SCF. We present such an example in the first section of the Online Appendix.

The comparison of our main result with the restrictive results on continuous implementation in Bayesian equilibrium is clear: the additional requirement of asking that all the Bayesian incentive constraints be met in the expanded type spaces –and not the continuity requirement *per se*– is responsible for those restrictive results. This final section closes with a brief discussion of generalizations and applications.

Nonsimple Type Spaces: Our results generalize to nonsimple type spaces. Continuity and SIRBIC still characterize continuous implementation up to level-K whenever we can distinguish all types in the planner's model by their z^{th}-order beliefs, for some $z \geq 1$. The argument for the necessity of continuity and SIRBIC remains the same. For sufficiency, we start by noting that the SCF defines a continuous direct mechanism $\hat{\mu}$ on $\times_{i \in I} Q_i^z(\hat{T})$ because each $\hat{T}_i$ is homeomorphic to $Q_i^z(\hat{T})$. We can then continuously extend $\hat{\mu}$ to the space of all z^{th}-order beliefs $\times_{i \in I} Q_i^z(T^*)$ – recall that T^* is the universal type space – by again applying Lemma 1. This continuous extension μ continuously implements up to level-K given truthful behavioral anchors (i.e., level-0 reports his z^{th}-order belief truthfully), following the same arguments as in the proof of Theorem 1.

In general type spaces, continuous implementation up to level-K entails, in addition to SIRBIC, a stronger necessary condition, viz., f must be continuous^p.[15] If types can be distinguished by their z^{th}-order beliefs, then continuity^p and continuity of the SCF are equivalent conditions. In general, however, continuity^p implies continuity. The two conditions, continuity^p and SIRBIC, are also sufficient for continuous implementation up to level-K in general type spaces. We begin by defining a mechanism $\hat{\mu}$ on $\times_{i \in I} Q_i(\hat{T})$ such that $\hat{\mu}(q) = f(t)$, where t is any type profile such that $q(t) = q$. The mechanism $\hat{\mu}$ is well-defined because if $q(t) = q(t')$, then $f(t) = f(t')$ since f is continuous^p. Moreover, $\hat{\mu}$ is continuous.[16] Then we can continuously extend $\hat{\mu}$ to the space of all belief hierarchies $\times_{i \in I} T_i^*$ using Lemma 1. Now this continuous extension μ continuously implements up to level-K given truthful behavioral anchors (i.e., level-0 reports his entire belief hierarchy truthfully), following the same arguments as in the

[15] The proof of Theorem 1 already contains the argument. There, we show that whenever f is continuously implementable up to level-K, then for any sequence $t^n \overset{p}{\to} t$, we must have $f(t^n)$ converge to $f(t)$.

[16] Pick any sequence $q^n \to q$. Let t^n and t in $\hat{T}$ be such that $q(t^n) = q^n$ and $q(t) = q$. Then $t^n \overset{p}{\to} t$. As f is continuous^p, we have $\hat{\mu}(q^n) = f(t^n) \to f(t) = \hat{\mu}(q)$.

proof of Theorem 1.

With SIRBIC and continuity^p characterizing continuous implementation up to level-K in all type spaces, Theorem 2 remains the same. Thus, regardless of the planner's model, the key difference between continuous strict equilibrium-implementation and continuous implementation up to level-K is that the former imposes an additional condition, i.e., the existence of extensions that satisfy Bayesian incentive compatibility.

Other Behavioral Anchors: dCSS show that SIRBIC is the key restriction on merely exact implementation up to level-K irrespective of the level-0 behavioral anchors. For atomless anchors (e.g., uniform-random play), they show that SIRBIC and continuity are sufficient for merely exact implementation up to level-K in environments with independent private values (IPV), i.e., when values are private and the type space is a common-prior payoff-type space with independent types (recall that such type spaces are simple). To prove that result, they construct an indirect mechanism in which players report both their types and real number from an interval. The key feature of that construction is that it separates the beliefs of level-1 from those of level-2 and above, inducing each level to report their types truthfully. A similar construction on the planner's model $\hat{T}$, followed by an extension of the mechanism to allow for reports in $(\Delta\Theta)^I \setminus \times_{i\in I} Q_i^1(\hat{T})$, can be used to prove that continuity and SIRBIC are sufficient for continuous implementation up to level-K given atomless anchors in most IPV settings (see Theorem 4 in the second section of the Online Appendix). Thus, in these cases too, SIRBIC and continuity remain the key conditions characterizing continuous implementation up to level-K.

Outside IPV environments, dCSS show that additional restrictions are necessary for merely exact implementation up to level-K when the level-0 anchors are type independent (e.g., uniform-random play). Specifically, a type separability condition that requires distinct types (i.e., those over which the SCF is responsive) have different preferences over constant lotteries becomes necessary (Theorem 5). In the third section of the Online Appendix we show in Theorem 6 that continuity, SIRBIC, and type separability are also sufficient for continuous implementation up to level-K given atomless (type-independent or otherwise) anchors in most non-IPV environments.

Applications: In the final section of the Online Appendix, we feature several well-known applications to showcase the permissiveness of continuous level-k implementation. First, in a large class of public decision problems, generalizing the private-values environments in d'Aspremont and Gerard-Varet (1979), we show that the SCF that extends their ex-post efficient rule satisfies continuity and strict incentive compatibility, making it implementable using the mechanisms we construct in Theorems 1 and 4 –the latter, for the particular case of private-values. Furthermore, allowing large fines, the SCF also satisfies the type separability condition, making it implementable using the mechanism in Theorem 6 as well. Second, in the bilateral trading setup with independent private values of Myerson and Satterthwaite (1983), although the second-best rule they propose is discontinuous and only weakly Bayesian incentive compatible, we are able to show, under monotone hazard rates, that a slight perturbation thereof is continuous and satisfies strict incentive compatibility, making this approximation continuously implementable up to level k, appealing to our mechanisms in Theorems 1 and 4. Finally, in the multidimensional type model of Jehiel *et al.* (2012), where locally robust implementation in their sense is impossible, we offer some permissive results for continuous level-k implementation, appealing to Theorem 6, even though this good news has to be more limited, as the behavioral anchors must depend nontrivially on the players' types. In particular, with type-independent anchors, there is a failure of the type separability condition, identified as necessary in Theorem 5.

A Appendix

Continuous Extension of a Continuous Mechanism:

Lemma 1. *Suppose $\hat{\mu} : \times_{i\in I}\hat{M}_i \to \Delta X$ is continuous and the message space $\hat{M}_i$ is a closed subset of some compact metric space M_i for all i. Then for each $i \in I$, there exists a correspondence $\omega_i : M_i \to \hat{M}_i$ with nonempty finite values and for each message $m_i \in M_i$, there exists a probability distribution ξ_{m_i} with full support on $\omega_i(m_i)$ such that $\mu : \times_{i\in I}M_i \to \Delta X$ extends $\hat{\mu}$ continuously, where μ is the mechanism that associates to any message profile $m \in \times_{i\in I}M_i$ the lottery that selects $\hat{\mu}(\hat{m})$ with probability $\times_{i\in I}\xi_{m_i}(\hat{m}_i)$, for all $\hat{m} \in \times_{i\in I}\omega_i(m_i)$.*

Proof. For each $i \in I$, define $Z_i = M_i \setminus \hat{M}_i$. Let $d : M_i \times M_i \to \mathbb{R}_+$ be the metric on M_i. Pick any $z_i \in Z_i$ and let $d(z_i, \hat{M}_i) = \inf\{d(z_i, \hat{m}_i) : \hat{m}_i \in \hat{M}_i\}$. Since Z_i is open (as $\hat{M}_i$ is closed by assumption), $d(z_i, \hat{M}_i) > 0$. Let $B(z_i, \frac{d(z_i,\hat{M}_i)}{4})$ be an open ball around z_i of radius $\frac{d(z_i,\hat{M}_i)}{4}$. Note that $B(z_i, \frac{d(z_i,\hat{M}_i)}{4}) \subset Z_i$. Now, $\left\{B(z_i, \frac{d(z_i,\hat{M}_i)}{4})\right\}_{z_i \in Z_i}$ is an open cover of Z_i. Since Z_i is a metric space, it is paracompact. Therefore, the open cover $\left\{B(z_i, \frac{d(z_i,\hat{M}_i)}{4})\right\}_{z_i \in Z_i}$ has a continuous locally finite partition of unity subordinate to it (see Theorem 2.90 in Aliprantis and Border (2006)). That is, there exists a family of functions $\{h_{z_i}\}_{z_i \in Z_i}$ from Z_i to $[0, 1]$ such that[17]

1. Each h_{z_i} is continuous.
2. Each $h_{z_i}(m_i) = 0$ if $m_i \in Z_i \setminus B(z_i, \frac{d(z_i,\hat{M}_i)}{4})$.
3. At each $m_i \in Z_i$, only finitely-many functions in the family $\{h_{z_i}\}_{z_i \in Z_i}$ are nonzero and $\sum_{z_i \in Z_i} h_{z_i}(m_i) = 1$.
4. Each $m_i \in Z_i$ has a neighborhood on which all but finitely-many functions in the family vanish.

For each $z_i \in Z_i$, let $\rho_i(z_i) \in \hat{M}_i$ be such that $d(z_i, \rho_i(z_i)) < \frac{5}{4} d(z_i, \hat{M}_i)$.

For each $i \in I$, define the correspondence $\omega_i : M_i \to \hat{M}_i$ as follows:

$$\omega_i(m_i) = \begin{cases} \{m_i\}, & \text{if } m_i \in \hat{M}_i \\ \{\rho_i(z_i) : z_i \in Z_i \text{ and } h_{z_i}(m_i) > 0\}, & \text{if } m_i \in Z_i. \end{cases}$$

Note that ω_i is finite-valued because of the third property of the collection $\{h_{z_i}\}_{z_i \in Z_i}$.

For each $m_i \in M_i$, define the probability distribution ξ_{m_i} over $\hat{M}_i$ as follows:

$$\xi_{m_i}(\hat{m}_i) = \begin{cases} 1, & \text{if } m_i \in \hat{M}_i \text{ and } \hat{m}_i = m_i \\ \sum_{z_i \in Z_i : \rho_i(z_i) = \hat{m}_i} h_{z_i}(m_i), & \text{if } m_i \in Z_i \text{ and } \hat{m}_i \in \omega_i(m_i) \\ 0, & \text{otherwise.} \end{cases}$$

[17] See Dugundji (1951) and Arens (1952) for a construction of such a family of functions. For example, taking $\mathbb{R}$ as a paracompact space, and $\cup_{z \in \mathbb{Z}}\{(z-1, z+1)\}$ as its open cover, and $h_z(x) = \min\{x - (z-1), z+1-x\}$ on $[z-1, z+1]$, and 0 otherwise. Then, for each $r \in \mathbb{R}$, let $h_r = h_{Int(r)}$. For each r, at most two of these functions, $h_{Int(r)}$ and either $h_{Int(r)-1}$ or $h_{Int(r)+1}$, do not vanish and their images add up to unity. Thus, each real number is covered by a finite number of open sets, each with a different weight, and the sum of these weights is always 1.

Thus, the support of ξ_{m_i} coincides with $\omega_i(m_i)$.

Now, define $\mu : \times_{i \in I} M_i \to \Delta X$ as follows:

$$\mu(m) = \sum_{\hat{m} \in \times_{i \in I} \omega_i(m_i)} \times_{i \in I} \xi_{m_i}(\hat{m}_i) \times \hat{\mu}(\hat{m}).$$

Since $\mu(m) = \hat{\mu}(m), \forall m \in \times_{i \in I} \hat{M}_i$, the mechanism μ is an extension of $\hat{\mu}$ to $\times_{i \in I} M_i$.

We now argue that μ is continuous. Let $(m^n)_{n \geq 1}$ be a sequence in $\times_{i \in I} M_i$ that converges to m. Pick any Borel subset A of X such that $\mu(m)(\partial A) = 0$. We argue that $\lim_{n \to \infty} \mu(m^n)(A) = \mu(m)(A)$. This is equivalent to proving that the sequence of probability measures $(\mu(m^n))_{n \geq 1}$ converges to $\mu(m)$ in the weak* topology.

Let's partition I into I_1, I_2 and I_3 such that

$$\begin{aligned} I_1 &= \{i \in I : m_i \text{ is in } Z_i\} \\ I_2 &= \{i \in I : m_i \text{ is in the interior of } \hat{M}_i\} \\ I_3 &= \{i \in I : m_i \text{ is on the boundary of } \hat{M}_i\}. \end{aligned}$$

Case 1. $i \in I_1$: Since $m_i \in Z_i$, there is a neighborhood $\mathcal{N}_i$ of m_i, with $\mathcal{N}_i \subseteq Z_i$, on which all but finitely-many functions in the family $\{h_{z_i}\}_{z_i \in Z_i}$ vanish. Let Z_i^* be the finite set of indices of the functions in this neighborhood that do not vanish. There exists n_i^* such that $m_i^n \in \mathcal{N}_i$ for all $n \geq n_i^*$. Therefore, if $n \geq n_i^*$, then $h_{z_i}(m_i^n) > 0 \implies z_i \in Z_i^*$, and so $\omega_i(m_i^n) \subseteq \{\rho_i(z_i) : z_i \in Z_i^*\}$.

Case 2. $i \in I_2$: Since m_i is in the interior of $\hat{M}_i$, there exists n_i^* such that $m_i^n \in \hat{M}_i$ for all $n \geq n_i^*$.

Case 3. $i \in I_3$: In this case, m_i is on the boundary of $\hat{M}_i$. For each m_i^n, pick any $\hat{m}_i^n \in \omega_i(m_i^n)$. We claim that the sequence $(\hat{m}_i^n)_{n \geq 1}$ converges to m_i in the weak* topology. The following two arguments are sufficient to establish this claim:

First, if there is any infinite subsequence $(m_i^{n_l})_{n_l \geq 1}$ such that $m_i^{n_l} \in \hat{M}_i, \forall n_l \geq 1$, then $\hat{m}_i^{n_l} = m_i^{n_l}, \forall n_l \geq 1$, and so the subsequence $(\hat{m}_i^{n_l})_{n_l \geq 1}$ converges to m_i.

Second, if there is any infinite subsequence $(m_i^{n_l})_{n_l \geq 1}$ such that $m_i^{n_l} \in Z_i, \forall n_l \geq 1$, then let $z_i^{n_l}$ be such that $\rho_i(z_i^{n_l}) = \hat{m}_i^{n_l}$ and $h_{z_i^{n_l}}(m_i^{n_l}) > 0$. Pick any $\epsilon > 0$ and consider the open ball $B(m_i, \frac{\epsilon}{3})$. Since $m_i^{n_l}$ converges to m_i, there exists n_i such that

$m_i^{n_l} \in B(m_i, \frac{\epsilon}{3})$ for all $n_l \geq n_i$. Hence, $m_i^{n_l} \in Z_i \cap B(m_i, \frac{\epsilon}{3})$ for all $n_l \geq n_i$. We argue that $d(m_i, \hat{m}_i^{n_l}) < \epsilon$ for all $n_l \geq n_i$. Note that

$$\begin{aligned} d(m_i, \hat{m}_i^{n_l}) &\leq d(m_i, m_i^{n_l}) + d(m_i^{n_l}, \hat{m}_i^{n_l}) \\ &\leq d(m_i, m_i^{n_l}) + d(m_i^{n_l}, z_i^{n_l}) + d(z_i^{n_l}, \hat{m}_i^{n_l}) \\ &< d(m_i, m_i^{n_l}) + d(m_i^{n_l}, z_i^{n_l}) + \frac{5}{4} d(z_i^{n_l}, \hat{M}_i). \end{aligned}$$

Since $h_{z_i^{n_l}}(m_i^{n_l}) > 0$, we have $d(m_i^{n_l}, z_i^{n_l}) < \frac{d(z_i^{n_l}, \hat{M}_i)}{4}$. Hence, $d(m_i, \hat{m}_i^{n_l}) < d(m_i, m_i^{n_l}) + \frac{6}{4} d(z_i^{n_l}, \hat{M}_i)$.

Next,

$$d(z_i^{n_l}, \hat{M}_i) \leq d(z_i^{n_l}, m_i) \leq d(z_i^{n_l}, m_i^{n_l}) + d(m_i^{n_l}, m_i) < \frac{d(z_i^{n_l}, \hat{M}_i)}{4} + d(m_i^{n_l}, m_i).$$

Therefore, $\frac{3}{4} d(z_i^{n_l}, \hat{M}_i)) < d(m_i^{n_l}, m_i)$. As a result,

$$d(m_i, \hat{m}_i^{n_l}) < d(m_i, m_i^{n_l}) + \frac{6}{4} d(z_i^{n_l}, \hat{M}_i) < 3 d(m_i, m_i^{n_l}) < \epsilon.$$

Hence, the subsequence $(\hat{m}_i^{n_l})_{n_l \geq 1}$ converges to m_i.

Now, by definition of $\mu(m^n)$, for any Borel $A \subseteq X$ such that $\mu(m)(\partial A) = 0$, we have

$$\mu(m^n)(A) = \sum_{\hat{m} \in \times_{i \in I} \omega_i(m_i^n)} \times_{i \in I} \xi_{m_i^n}(\hat{m}_i) \times \hat{\mu}(\hat{m})(A).$$

Consider any $n \geq n^* = \max\{n_i^* : i \in I_1 \cup I_2\}$. Then $m_i^n \in \hat{M}_i, \forall i \in I_2$. Hence,

$$\mu(m^n)(A) = \sum_{(\hat{m}_i)_{i \in I_1 \cup I_3} \in \times_{i \in I_1 \cup I_3} \omega_i(m_i^n)} \times_{i \in I_1 \cup I_3} \xi_{m_i^n}(\hat{m}_i) \times \hat{\mu}((\hat{m}_i)_{i \in I_1 \cup I_3}, (m_i^n)_{i \in I_2})(A). \tag{5}$$

Pick any $(\hat{m}_i)_{i \in I_3} \in \times_{i \in I_3} \omega_i(m_i^n)$, and define

$$Y^n((\hat{m}_i)_{i \in I_3}) = \sum_{(\hat{m}_i)_{i \in I_1} \in \times_{i \in I_1} \omega_i(m_i^n)} \times_{i \in I_1} \xi_{m_i^n}(\hat{m}_i) \times \hat{\mu}((\hat{m}_i)_{i \in I_1}, (\hat{m}_i)_{i \in I_3}, (m_i^n)_{i \in I_2})(A).$$

Then it follows from (5) that

$$\mu(m^n)(A) = \sum_{(\hat{m}_i)_{i\in I_3}\in \times_{i\in I_3}\omega_i(m_i^n)} \times_{i\in I_3}\xi_{m_i^n}(\hat{m}_i) Y^n((\hat{m}_i)_{i\in I_3}).$$

Since $\times_{i\in I_3}\omega_i(m_i^n)$ is a finite set, we can find $(\hat{m}_i^{1n})_{i\in I_3} \in \times_{i\in I_3}\omega_i(m_i^n)$ such that $Y^n((\hat{m}_i^{1n})_{i\in I_3}) \geq Y^n((\hat{m}_i)_{i\in I_3}), \forall (\hat{m}_i)_{i\in I_3} \in \times_{i\in I_3}\omega_i(m_i^n)$. Similarly, we can find $(\hat{m}_i^{2n})_{i\in I_3} \in \times_{i\in I_3}\omega_i(m_i^n)$ such that $Y^n((\hat{m}_i^{2n})_{i\in I_3}) \leq Y^n((\hat{m}_i)_{i\in I_3}), \forall (\hat{m}_i)_{i\in I_3} \in \times_{i\in I_3}\omega_i(m_i^n)$. Hence, $Y^n((\hat{m}_i^{1n})_{i\in I_3}) \geq \mu(m^n)(A) \geq Y^n((\hat{m}_i^{2n})_{i\in I_3})$. We argue that $\lim_{n\to\infty} Y^n((\hat{m}_i^{1n})_{i\in I_3}) = \lim_{n\to\infty} Y^n((\hat{m}_i^{2n})_{i\in I_3}) = \mu(m)(A)$, which implies that $\lim_{n\to\infty}\mu(m^n)(A) \to \mu(m)(A)$.

As $n \geq n^*$, we have $m_i^n \in \mathcal{N}_i \subseteq Z_i, \forall i \in I_1$. Then, as argued in Case 1 above, $\omega_i(m_i^n) \subseteq \{\rho_i(z_i) : z_i \in Z_i^*\}, \forall i \in I_1$. Hence,

$$Y^n((\hat{m}_i^{1n})_{i\in I_3}) = \sum_{(\hat{m}_i)_{i\in I_1}\in \times_{i\in I_1}\{\rho_i(z_i):z_i\in Z_i^*\}} \times_{i\in I_1}\xi_{m_i^n}(\hat{m}_i)\times\hat{\mu}((\hat{m}_i)_{i\in I_1}, (\hat{m}_i^{1n})_{i\in I_3}, (m_i^n)_{i\in I_2})(A).$$

Take any $i \in I_1$ and $\hat{m}_i \in \{\rho_i(z_i) : z_i \in Z_i^*\}$. Since $m_i^n \in \mathcal{N}_i$, we have $\xi_{m_i^n}(\hat{m}_i) = \sum_{z_i\in Z_i^*:\rho_i(z_i)=\hat{m}_i} h_{z_i}(m_i^n)$. As each h_{z_i} is continuous,

$$\lim_{n\to\infty}\xi_{m_i^n}(\hat{m}_i) = \sum_{z_i\in Z_i^*:\rho_i(z_i)=\hat{m}_i} h_{z_i}(m_i) = \xi_{m_i}(\hat{m}_i).$$

It follows from the arguments made in Case 3 above that for all $i \in I_3$, $\hat{m}_i^{1n}$ converges to m_i. Hence, as $\hat{\mu}$ is continuous, we obtain

$$\begin{aligned}\lim_{n\to\infty} Y^n((\hat{m}_i^{1n})_{i\in I_3}) &= \sum_{(\hat{m}_i)_{i\in I_1}\in \times_{i\in I_1}\{\rho_i(z_i):z_i\in Z_i^*\}} \times_{i\in I_1}\xi_{m_i}(\hat{m}_i) \times \hat{\mu}((\hat{m}_i)_{i\in I_1}, (m_i)_{i\in I_2\cup I_3})(A)\\ &= \mu(m)(A).\end{aligned}$$

A similar argument shows that $\lim_{n\to\infty} Y^n((\hat{m}_i^{2n})_{i\in I_3}) = \mu(m)(A)$. Therefore, μ is continuous. □

Technical Lemmata: We now present a set of four technical lemmata that prove some general properties of level-k behavior. These results do not restrict to truthful behavioral anchors. Indeed, we let $\alpha_i^{\mu,\mathcal{T}}$ be the level-0 behavioral anchors of individual i, which captures what other individuals think would be the gut reaction of individual

i when she is confronted to play the mechanism μ on the type space $\mathcal{T}$. Profiles of such anchors will be denoted $\alpha^{\mu,\mathcal{T}} = (\alpha_i^{\mu,\mathcal{T}})_{i\in I}$.

Given a type space $\mathcal{T}$, mechanism μ, and behavioral anchors $\alpha^{\mu,\mathcal{T}}$, we let $S_i^k(\alpha^{\mu,\mathcal{T}})$ denote the set of level-k consistent strategies of individual i, starting with level-1 consistent strategies that support only those messages that are best responses to others playing $\alpha_{-i}^{\mu,\mathcal{T}}$. Like before, we define the correspondence $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}})$ that associates to each type of individual i, his best response messages against all conjectures in $\Delta(\Theta\times T_{-i}\times M_{-i})$ 'consistent' with the behavior of level-$(k-1)$ opponents (with $\alpha_{-i}^{\mu,\mathcal{T}}$ at level-0). We are now ready to present the four lemmata.

Lemma 2. *Suppose the mechanism μ is continuous. Then for any type space $\mathcal{T}$, individual i, and $\sigma_{-i} : T_{-i} \to \Delta M_{-i}$, the best-response correspondence $BR_i^{\sigma_{-i}} : T_i \to M_i$ is nonempty and compact valued, and admits a measurable selector. Additionally, if σ_{-i} is continuous, then $BR_i^{\sigma_{-i}}$ is also upper hemicontinuous.*

Proof. Fix the type space $\mathcal{T}$, individual i, and $\sigma_{-i} : T_{-i} \to \Delta M_{-i}$ in the mechanism μ. For any $m_i \in M_i$ and $t_i \in T_i$, define

$$\begin{aligned} W_i(m_i, t_i) &= \int_{\Theta\times T_{-i}} U_i(\mu(m_i, \sigma_{-i}(t_{-i})), \theta) d\pi_i(t_i) \\ &= \int_{\Theta\times T_{-i}} \int_{M_{-i}} U_i(\mu(m_i, m_{-i}), \theta) d\sigma_{-i}(t_{-i}) d\pi_i(t_i). \end{aligned}$$

We argue that W_i is a Carathéodory function such that for each t_i, $W_i(\cdot, t_i) : M_i \to \mathbb{R}$ is continuous whereas for each m_i, $W_i(m_i, \cdot) : T_i \to \mathbb{R}$ is measurable.

Let $(m_i^n)_{n\geq 1}$ be a sequence such that $m_i^n \to m_i$. Recall that U_i is continuous and bounded and μ is continuous. Therefore, $U_i(\mu(\cdot), \cdot)$, as a function of message profiles in M and state in Θ, is continuous and bounded over a compact metric space. Hence, it is uniformly continuous. Therefore, for every $\epsilon > 0$, there exists n' such that if $n \geq n'$, then $|U_i(\mu(m_i^n, m_{-i}), \theta) - U_i(\mu(m_i, m_{-i}), \theta)| < \epsilon$, for all $(m_{-i}, \theta) \in M_{-i} \times \Theta$. Therefore, for all $n \geq n'$, we have

$$\begin{aligned} &|W_i(m_i^n, t_i) - W_i(m_i, t_i)| \\ &\leq \int_{\Theta\times T_{-i}} \int_{M_{-i}} |U_i(\mu(m_i^n, m_{-i}), \theta) - U_i(\mu(m_i, m_{-i}), \theta)| d\sigma_{-i}(t_{-i}) d\pi_i(t_i) < \epsilon. \end{aligned}$$

So $W_i(\cdot, t_i)$ is continuous in m_i.

To argue that $W_i(m_i, \cdot)$ is measurable, consider the mapping $h_i : \Theta \times T_{-i} \to \mathbb{R}$ where

$$h_i(\theta, t_{-i}) = \int_{M_{-i}} U_i(\mu(m_i, m_{-i}), \theta) d\sigma_{-i}(t_{-i}).$$

Using similar arguments as for the case of $W_i(\cdot, t_i)$, we can argue that $h_i(\cdot, t_{-i})$ is continuous in θ. For a fixed θ, the function $h_i(\theta, \cdot)$ is a composition of the mapping $\tilde{\eta} \to \int_{M_{-i}} U_i(\mu(m_i, m_{-i}), \theta) d\tilde{\eta}$, where $\tilde{\eta} \in \Delta M_{-i}$, and $\sigma_{-i} : T_{-i} \to \Delta M_{-i}$. As U_i is continuous and bounded, and μ is continuous, it follows from the definition of weak convergence that the mapping $\tilde{\eta} \to \int_{M_{-i}} U_i(\mu(m_i, m_{-i}), \theta) d\tilde{\eta}$ is continuous. Since σ_{-i} is a measurable, $h_i(\theta, \cdot)$ is measurable in t_{-i}. (Notice that if σ_{-i} were continuous, then $h_i(\theta, \cdot)$ would be continuous in t_{-i}. We use this fact below.) Hence, h_i is a Carathéodory function, which implies that h_i is jointly measurable (see Lemma 4.51 in Aliprantis and Border (2006)). Then the mapping $\tilde{\gamma} \to \int_{\Theta \times T_{-i}} h_i(\theta, t_{-i}) d\tilde{\gamma}$, where $\tilde{\gamma} \in \Delta(\Theta \times T_{-i})$ is (Borel) measurable (see Theorem 15.13 in Aliprantis and Border (2006)). Now, $W_i(m_i, \cdot)$ is a composition of the mapping $\tilde{\gamma} \to \int_{\Theta \times T_{-i}} h_i(\theta, t_{-i}) d\tilde{\gamma}$ and $\pi_i : T_i \to \Delta(\Theta \times T_{-i})$. Since π_i is measurable, we conclude that $W_i(m_i, \cdot)$ is measurable in t_i.

Now that we have established that W_i is a Carathéodory function, and since $BR_i^{\sigma_{-i}}(t_i) = \arg\max_{m_i \in M_i} W_i(m_i, t_i)$ for all t_i, we can use the Measurable Maximum Theorem (see Theorem 18.19 in Aliprantis and Border (2006)) to claim that $BR_i^{\sigma_{-i}}$ is nonempty and compact valued, and admits a measurable selector.

To prove the additional statement in the lemma, suppose σ_{-i} is continuous. We then argue that W_i is a continuous function.

The first step is to argue that h_i is continuous. Let $(\theta^n)_{n \geq 1}$ and $(t_{-i}^n)_{n \geq 1}$ be two sequences such that $\theta^n \to \theta$ and $t_{-i}^n \to t_{-i}$. As mentioned above, if σ_{-i} is continuous, then $h_i(\theta, \cdot)$ is continuous in t_{-i}. Hence, $h_i(\theta, t_{-i}^n)$ converges to $h_i(\theta, t_{-i})$. That is, for every $\epsilon > 0$, there exists n_1 such that if $n \geq n_1$, then $|h_i(\theta, t_{-i}^n) - h_i(\theta, t_{-i})| < \frac{\epsilon}{2}$. Using the fact that $U_i(\mu(\cdot), \cdot)$ is uniformly continuous, we know that for every $\epsilon > 0$, there exists n_2 such that if $n \geq n_2$, then $|U_i(\mu(m_i, m_{-i}), \theta^n) - U_i(\mu(m_i, m_{-i}), \theta)| < \frac{\epsilon}{2}$, for all $(m_i, m_{-i}) \in M$. Therefore, for all $n \geq n_2$, we have

$$\begin{aligned}
&|h_i(\theta^n, t^n_{-i}) - h_i(\theta, t^n_{-i})| \\
&\leq \int_{M_{-i}} |U_i(\mu(m_i, m_{-i}), \theta^n) - U_i(\mu(m_i, m_{-i}), \theta)| d\sigma_{-i}(t^n_{-i}) < \frac{\epsilon}{2}.
\end{aligned}$$

Hence, for all $n \geq \max\{n_1, n_2\}$, we have

$$|h_i(\theta^n, t^n_{-i}) - h_i(\theta, t_{-i})| \leq |h_i(\theta^n, t^n_{-i}) - h_i(\theta, t^n_{-i})| + |h_i(\theta, t^n_{-i}) - h_i(\theta, t_{-i})| < \epsilon.$$

Therefore, h_i is continuous.

The final step is to show that W_i itself is continuous. Let $(m^n_i)_{n\geq 1}$ and $(t^n_i)_{n\geq 1}$ be two sequences such that $m^n_i \to m_i$ and $t^n_i \to t_i$. Since h_i is continuous and bounded (as U_i is bounded), it follows from the definition of weak convergence that $W_i(m_i, t^n_i)$ converges to $W_i(m_i, t_i)$. That is, for every $\epsilon > 0$, there exists n_1 such that if $n \geq n_1$, then $|W_i(m_i, t^n_i) - W_i(m_i, t_i)| < \frac{\epsilon}{2}$. Again, using the fact that $U_i(\mu(\cdot), \cdot)$ is uniformly continuous, we know that for every $\epsilon > 0$, there exists n_2 such that if $n \geq n_2$, then $|U_i(\mu(m^n_i, m_{-i}), \theta) - U_i(\mu(m_i, m_{-i}), \theta)| < \frac{\epsilon}{2}$, for all $(m_{-i}, \theta) \in M_{-i} \times \Theta$. Therefore, for all $n \geq n_2$, we have

$$\begin{aligned}
&|W_i(m^n_i, t^n_i) - W_i(m_i, t^n_i)| \\
&\leq \int_{\Theta \times T_{-i}} \int_{M_{-i}} |U_i(\mu(m^n_i, m_{-i}), \theta) - U_i(\mu(m_i, m_{-i}), \theta)| d\sigma_{-i}(t_{-i}) d\pi_i(t^n_i) < \frac{\epsilon}{2}.
\end{aligned}$$

Hence, for all $n \geq \max\{n_1, n_2\}$, we have

$$|W_i(m^n_i, t^n_i) - W_i(m_i, t_i)| \leq |W_i(m^n_i, t^n_i) - W_i(m_i, t^n_i)| + |W_i(m_i, t^n_i) - W_i(m_i, t_i)| < \epsilon.$$

Therefore, W_i is continuous. It follows from the Berge's Maximum Theorem that $BR_i^{\sigma_{-i}}$ is upper hemicontinuous. □

Lemma 3. *Suppose the mechanism μ is continuous. Consider any type space $\mathcal{T}$ and behavioral anchors $\alpha^{\mu,\mathcal{T}}$ such that $\alpha_i^{\mu,\mathcal{T}}$ is continuous for all i. Then the correspondence $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}})$ is upper hemicontinuous for all i and $k \geq 1$.*

Proof. We argue by induction that $Gr(\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}}))$, i.e., the graph of $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}})$, is closed for all i and $k \geq 1$. Since M_i is compact, this implies that $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}})$ is upper hemicontinuous for all i and $k \geq 1$.

As $\alpha_{-i}^{\mu,\mathcal{T}}$ is continuous, $\Sigma_i^1(\cdot|\alpha^{\mu,\mathcal{T}})$ is upper hemicontinuous and compact valued (Lemma 2). Hence, $Gr(\Sigma_i^1(\cdot|\alpha^{\mu,\mathcal{T}}))$ is closed for all i.

Next, consider $k > 1$, and suppose $Gr(\Sigma_i^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$ is closed for all i. Pick any individual i and consider sequences $(t_i^n)_{n\geq 1}$ and $(m_i^n)_{n\geq 1}$ such that $t_i^n \to t_i$, $m_i^n \to m_i$, and $m_i^n \in \Sigma_i^k(t_i^n|\alpha^{\mu,\mathcal{T}})$ for all $n \geq 1$. Since $m_i^n \in \Sigma_i^k(t_i^n|\alpha^{\mu,\mathcal{T}})$, there exists $\gamma^n \in \Delta(\Theta \times T_{-i} \times M_{-i})$, such that *(a)* the marginal of γ^n on $\Theta \times T_{-i}$ equals $\pi_i(t_i^n)$, *(b)* the marginal of γ^n on $T_{-i} \times M_{-i}$ supports a subset of $\times_{j\neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$, and

$$m_i^n \in \arg\max_{m_i' \in M_i} \int_{\Theta\times T_{-i}\times M_{-i}} U_i(\mu(m_i', m_{-i}), \theta) d\gamma^n.$$

Since $\Theta \times T_{-i} M_{-i}$ is a compact metric space, so is $\Delta(\Theta \times T_{-i} \times M_{-i})$. Hence, the sequence $(\gamma^n)_{n\geq 1}$ has a convergent subsequence $(\gamma^{n_l})_{n_l \geq 1}$ that converges to say γ in the weak* topology.

Since $\text{marg}_{\Theta\times T_{-i}}\gamma^{n_l} = \pi_i(t_i^{n_l}) \to \pi_i(t_i)$ and $\text{marg}_{\Theta\times T_{-i}}\gamma^{n_l} \to \text{marg}_{\Theta\times T_{-i}}\gamma$, we have that $\text{marg}_{\Theta\times T_{-i}}\gamma = \pi_i(t_i)$.

By the induction hypothesis, $Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$ is closed for all $j \neq i$. Hence, $\Theta \times_{j\neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$ is closed. The fact that γ^{n_l} converges to γ in the weak* topology implies that

$$\gamma\left(\Theta \times_{j\neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))\right) \geq \limsup_{n_l} \gamma^{n_l}\left(\Theta \times_{j\neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))\right) = 1.$$

Therefore, the marginal of γ on $T_{-i} \times M_{-i}$ supports a subset of $\times_{j\neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$.

For each $\hat{m}_i \in M_i$ and $\hat{\gamma} \in \Delta(\Theta \times T_{-i} \times M_{-i})$, define

$$V_i(\hat{m}_i, \hat{\gamma}) = \int_{\Theta\times T_{-i}\times M_{-i}} U_i(\mu(\hat{m}_i, m_{-i}), \theta) d\hat{\gamma}.$$

We argue that V_i is continuous. Let $(\hat{m}_i^n)_{n\geq 1}$ and $(\hat{\gamma}^n)_{n\geq 1}$ be two sequences such that $\hat{m}_i^n \to \hat{m}_i$ and $\hat{\gamma}^n \to \hat{\gamma}$. Since U_i is continuous and bounded and μ is continuous, it follows from the definition of weak convergence that $V_i(\hat{m}_i, \hat{\gamma}^n)$ converges to $V_i(\hat{m}_i, \hat{\gamma})$. That is, for every $\epsilon > 0$, there exists n_1 such that if $n \geq n_1$, then $|V_i(\hat{m}_i, \hat{\gamma}^n) -$

$V_i(\hat{m}_i, \hat{\gamma})| < \frac{\epsilon}{2}$.

Since $U_i(\mu(\cdot), \cdot)$, as a function of message profiles in M and state in Θ, is a continuous function over a compact metric space, it is uniformly continuous. Therefore, for every $\epsilon > 0$, there exists n_2 such that if $n \geq n_2$, then $|U_i(\mu(\hat{m}_i^n, m_{-i}), \theta) - U_i(\mu(\hat{m}_i, m_{-i}), \theta)| < \frac{\epsilon}{2}$, for all $(m_{-i}, \theta) \in M_{-i} \times \Theta$. Therefore, for all $n \geq n_2$, we have

$$\begin{aligned} &|V_i(\hat{m}_i^n, \hat{\gamma}^n) - V_i(\hat{m}_i, \hat{\gamma}^n)| \\ &\leq \int_{\Theta \times T_{-i} \times M_{-i}} |U_i(\mu(\hat{m}_i^n, m_{-i}), \theta) - U_i(\mu(\hat{m}_i, m_{-i}), \theta)| d\hat{\gamma}^n < \frac{\epsilon}{2}. \end{aligned}$$

Hence, for all $n \geq \max\{n_1, n_2\}$, we have

$$|V_i(\hat{m}_i^n, \hat{\gamma}^n) - V_i(\hat{m}_i, \hat{\gamma})| \leq |V_i(\hat{m}_i^n, \hat{\gamma}^n) - V_i(\hat{m}_i, \hat{\gamma}^n)| + |V_i(\hat{m}_i, \hat{\gamma}^n) - V_i(\hat{m}_i, \hat{\gamma})| < \epsilon.$$

Therefore, V_i is continuous. It follows from the Berge's Maximum Theorem that the correspondence $\hat{\gamma} \to \arg\max_{\hat{m}_i \in M_i} V_i(\hat{m}_i, \hat{\gamma})$ is upper hemicontinuous and compact valued.

The subsequences $(m_i^{n_l})_{n_l \geq 1}$ and $(\gamma^{n_l})_{n_l \geq 1}$ are such that $m_i^{n_l} \in \arg\max_{\hat{m}_i \in M_i} V_i(\hat{m}_i, \gamma^{n_l})$. So we must have $m_i \in \arg\max_{\hat{m}_i \in M_i} V_i(\hat{m}_i, \gamma)$. We thus conclude that $m_i \in \Sigma_i^k(t_i | \alpha^{\mu, \mathcal{T}})$, and so $Gr(\Sigma_i^k(\cdot | \alpha^{\mu, \mathcal{T}}))$ is closed. □

Lemma 4. *Suppose the mechanism μ continuous. Consider any type space $\mathcal{T}$ and behavioral anchors $\alpha^{\mu, \mathcal{T}}$ such that $\alpha_i^{\mu, \mathcal{T}}$ is continuous for all i. Then for any depth of reasoning $k \geq 1$, individual i, strategy $\sigma_i \in S_i^k(\alpha^{\mu, \mathcal{T}})$, and type $t_i \in T_i$, if m_i is in the support of $\sigma_i(t_i)$, then $m_i \in \Sigma_i^k(t_i | \alpha^{\mu, \mathcal{T}})$.*

Proof. We prove this by induction on k. Consider $\sigma_i \in S_i^1(\alpha^{\mu, \mathcal{T}})$. Then each m_i in the support of $\sigma_i(t_i)$ is an element of $BR_i^{\alpha_{-i}^{\mu, \mathcal{T}}}(t_i) = \Sigma_i^1(t_i | \alpha^{\mu, \mathcal{T}})$.

Next, consider $k > 1$ and suppose the statement is true for all $\sigma_i \in S_i^{k-1}(\alpha^{\mu, \mathcal{T}})$ and all i. Pick any $\sigma_i \in S_i^k(\alpha^{\mu, \mathcal{T}})$. Then there exists $\sigma_{-i} \in S_{-i}^{k-1}(\alpha^{\mu, \mathcal{T}})$ such that every m_i in the support of $\sigma_i(t_i)$ is an element of $BR_i^{\sigma_{-i}}(t_i)$. We can find a unique $\gamma \in \Delta(\Theta \times T_{-i} \times M_{-i})$ such that for all measurable $E \subseteq \Theta \times T_{-i}$ and $F \subseteq M_{-i}$,

$$\gamma(E \times F) = \int_{\Theta \times T_{-i}} \chi(E) \sigma_{-i}(t_{-i})(F) d\pi_i(t_i),$$

where $\chi(E)$ is the indicator function on E. Then the marginal distribution of γ on $\Theta \times T_{-i}$ is equal to $\pi_i(t_i)$.

Let (t_{-i}, m_{-i}) be in the support of the marginal of γ on $T_{-i} \times M_{-i}$. By the definition of the support, the marginal distribution of γ on $T_{-i} \times M_{-i}$ assigns a positive probability to every open neighborhood of (t_{-i}, m_{-i}). Hence, we can find a sequence $(t^n_{-i}, m^n_{-i})_{n\geq 1}$ converging to (t_{-i}, m_{-i}) such that m^n_j is in the support of $\sigma_j(t^n_j)$ for all $j \neq i$ and $n \geq 1$. By the induction hypothesis, $m^n_{-i} \in \times_{j\neq i}\Sigma^{k-1}_j(t^n_j|\alpha^{\mu,\mathcal{T}})$ for all n. Thus, $(t^n_{-i}, m^n_{-i}) \in \times_{j\neq i}Gr(\Sigma^{k-1}_j(\cdot|\alpha^{\mu,\mathcal{T}}))$ for all n. As argued in Lemma 3, $Gr(\Sigma^{k-1}_j(\cdot|\alpha^{\mu,\mathcal{T}}))$ is closed for all $j \neq i$. Hence, $(t_{-i}, m_{-i}) \in \times_{j\neq i}Gr(\Sigma^{k-1}_j(\cdot|\alpha^{\mu,\mathcal{T}}))$. Thus, the marginal of γ on $T_{-i} \times M_{-i}$ supports a subset of $\times_{j\neq i}Gr(\Sigma^{k-1}_j(\cdot|\alpha^{\mu,\mathcal{T}}))$.

Finally, since

$$\int_{\Theta\times T_{-i}}\int_{M_{-i}} U_i(\mu(m'_i, m_{-i}), \theta)d\sigma_{-i}(t_{-i})d\pi_i(t_i) = \int_{\Theta\times T_{-i}\times M_{-i}} U_i(\mu(m'_i, m_{-i}), \theta)d\gamma,$$

for all m'_i, if $m_i \in BR^{\sigma_{-i}}_i(t_i)$, then m_i is a best response against the conjecture γ. Hence, if m_i is in the support of $\sigma_i(t_i)$, then $m_i \in \Sigma^k_i(t_i|\alpha^{\mu,\mathcal{T}})$. □

Lemma 5. *Suppose the mechanism μ is continuous. Consider any type space $\mathcal{T}$ and behavioral anchors $\alpha^{\mu,\mathcal{T}}$ such that $\alpha^{\mu,\mathcal{T}}_i$ is continuous^p for all i. If $q_i(t_i) = q_i(t'_i)$, then $\Sigma^k_i(t_i|\alpha^{\mu,\mathcal{T}}) = \Sigma^k_i(t'_i|\alpha^{\mu,\mathcal{T}})$ for all $k \geq 1$.*

Proof. We argue by induction on k. First consider $\Sigma^1_i(\cdot|\alpha^{\mu,\mathcal{T}})$. Define $\hat{\sigma}_{-i} : Q_{-i}(T) \to \Delta M_{-i}$ as $\hat{\sigma}_{-i}(q_{-i}) = \alpha^{\mu,\mathcal{T}}_{-i}(t_{-i})$ for any t_{-i} such that $q_{-i}(t_{-i}) = q_{-i}$. Notice that $\hat{\sigma}_{-i}$ is a well-defined function since if $q_{-i}(t_{-i}) = q_{-i}(t'_{-i})$, then $\alpha^{\mu,\mathcal{T}}_{-i}(t_{-i}) = \alpha^{\mu,\mathcal{T}}_{-i}(t'_{-i})$ because $\alpha^{\mu,\mathcal{T}}_{-i}$ is continuous^p. Thus defined, $\hat{\sigma}_{-i}$ is in fact continuous (and hence, measurable). To see this, pick any sequence $(q^n_{-i})_{n\geq 1}$ that converges to q_{-i}. Let t^n_{-i} and t_{-i} be such that $q_{-i}(t^n_{-i}) = q^n_{-i}$ and $q_{-i}(t_{-i}) = q_{-i}$. Thus, $q_{-i}(t^n_{-i})$ converges to $q_{-i}(t_{-i})$ or equivalently, $t^n_{-i} \xrightarrow{p} t_{-i}$. Since $\alpha^{\mu,\mathcal{T}}_{-i}$ is continuous^p, $\hat{\sigma}_{-i}(q^n_{-i}) = \alpha^{\mu,\mathcal{T}}_{-i}(t^n_{-i})$ converges to $\hat{\sigma}_{-i}(q_{-i}) = \alpha^{\mu,\mathcal{T}}_{-i}(t_{-i})$.

Recall that q_{-i} is a belief-preserving morphism from T_{-i} to T^*_{-i}. Therefore, for any m_i and t_i, we have

$$\int_{\Theta\times T_{-i}}\int_{M_{-i}} U_i(\mu(m_i, m_{-i}), \theta)d\alpha^{\mu,\mathcal{T}}_{-i}(t_{-i})d\pi_i(t_i)$$

$$= \int_{\Theta \times Q_{-i}(T)} \int_{M_{-i}} U_i(\mu(m_i, m_{-i}), \theta) d\hat{\sigma}_{-i}(q_{-i}) d\pi_i^*(q_i(t_i)).$$

Thus, if t_i and t_i' are such that $q_i(t_i) = q_i(t_i')$, then $BR_i^{\alpha_{-i}^{\mu,\mathcal{T}}}(t_i) = BR_i^{\alpha_{-i}^{\mu,\mathcal{T}}}(t_i')$. So $\Sigma_i^1(t_i|\alpha^{\mu,\mathcal{T}}) = \Sigma_i^1(t_i'|\alpha^{\mu,\mathcal{T}})$.

Next, consider $k > 1$ and suppose the statement is true for $k-1$. Pick $m_i \in \Sigma_i^k(t_i|\alpha^{\mu,\mathcal{T}})$. Then

$$m_i \in \arg\max_{m_i' \in M_i} \int_{\Theta \times T_{-i} \times M_{-i}} U_i(\mu(m_i', m_{-i}), \theta) d\gamma$$

for some conjecture $\gamma \in \Delta(\Theta \times T_{-i} \times M_{-i})$ such that *(a)* the distribution $\pi_i(t_i)$ coincides with the marginal distribution of γ on $\Theta \times T_{-i}$, and *(b)* the marginal distribution of γ on $T_{-i} \times M_{-i}$ supports a subset of $\times_{j \neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$.

Consider the continuous mapping $h : \Theta \times T_{-i} \times M_{-i} \to \Theta \times Q_{-i}(T)$ such that $h(\theta, t_{-i}, m_{-i}) = (\theta, q_{-i}(t_{-i}))$. As Θ, T_{-i} and M_{-i} are compact metric spaces, there exists a version of regular conditional probabilities $\gamma_{(\theta,q_{-i})}$ such that for all $m_i' \in M_i$,

$$\begin{aligned}
&\int_{\Theta \times T_{-i} \times M_{-i}} U_i(\mu(m_i', m_{-i}), \theta) d\gamma \\
&= \int_{\Theta \times Q_{-i}(T)} \int_{\{(\theta, t_{-i}) : q_{-i}(t_{-i}) = q_{-i}\} \times M_{-i}} U_i(\mu(m_i', m_{-i}), \theta) d\gamma_{(\theta,q_{-i})} d\pi_i^*(q_i(t_i)) \\
&= \int_{\Theta \times Q_{-i}(T)} \int_{M_{-i}} U_i(\mu(m_i', m_{-i}), \theta) d\text{marg}_{M_{-i}} \gamma_{(\theta,q_{-i})} d\pi_i^*(q_i(t_i)),
\end{aligned}$$

where $\text{marg}_{M_{-i}} \gamma_{(\theta,q_{-i})}$ is the marginal distribution of $\gamma_{(\theta,q_{-i})}$ on M_{-i}.

Pick any t_i' such that $q_i(t_i') = q_i(t_i)$. We can find a unique $\gamma' \in \Delta(\Theta \times T_{-i} \times M_{-i})$ such that for all measurable $E \subseteq \Theta \times T_{-i}$ and $F \subseteq M_{-i}$,

$$\gamma'(E \times F) = \int_{\Theta \times T_{-i}} \chi(E) \text{marg}_{M_{-i}} \gamma_{(\theta, q_{-i}(t_{-i}))}(F) d\pi_i(t_i'),$$

Then the marginal distribution of γ' on $\Theta \times T_{-i}$ is equal to $\pi_i(t_i')$.

Let (t_{-i}, m_{-i}) be in the support of the marginal of γ' on $T_{-i} \times M_{-i}$. For each $n \geq 1$, let $B_{\frac{1}{n}}(q_{-i}(t_{-i}))$ be the open ball of radius $\frac{1}{n}$ around $q_{-i}(t_{-i})$. The inverse image $q_{-i}^{-1}(B_{\frac{1}{n}}(q_{-i}(t_{-i})))$ is an open subset of T_{-i} containing t_{-i}. Let $\hat{B}_{\frac{1}{n}}(m_{-i})$ be the open ball of radius $\frac{1}{n}$ around m_{-i}. By the definition of support, the marginal distribution of γ' on $T_{-i} \times M_{-i}$ assigns a positive probability to $q_{-i}^{-1}(B_{\frac{1}{n}}(q_{-i}(t_{-i}))) \times \hat{B}_{\frac{1}{n}}(m_{-i})$. Hence, $\gamma'(\Theta \times q_{-i}^{-1}(B_{\frac{1}{n}}(q_{-i}(t_{-i}))) \times \hat{B}_{\frac{1}{n}}(m_{-i})) > 0$. But

$$\begin{aligned}
&\gamma'(\Theta \times q_{-i}^{-1}(B_{\frac{1}{n}}(q_{-i}(t_{-i}))) \times \hat{B}_{\frac{1}{n}}(m_{-i})) \\
&= \int_{\Theta \times T_{-i}} \chi(\Theta \times q_{-i}^{-1}(B_{\frac{1}{n}}(q_{-i}(t_{-i}))))\mathrm{marg}_{M_{-i}}\gamma_{(\theta, q_{-i}(t'_{-i}))}(\hat{B}_{\frac{1}{n}}(m_{-i}))d\pi_i(t'_i) \\
&= \int_{\Theta \times Q_{-i}(T)} \chi(\Theta \times B_{\frac{1}{n}}(q_{-i}(t_{-i})))\mathrm{marg}_{M_{-i}}\gamma_{(\theta, q_{-i})}(\hat{B}_{\frac{1}{n}}(m_{-i}))d\pi_i^*(q_i(t'_i)) \\
&= \int_{\Theta \times Q_{-i}(T)} \chi(\Theta \times B_{\frac{1}{n}}(q_{-i}(t_{-i})))\mathrm{marg}_{M_{-i}}\gamma_{(\theta, q_{-i})}(\hat{B}_{\frac{1}{n}}(m_{-i}))d\pi_i^*(q_i(t_i)) \\
&= \int_{\Theta \times Q_{-i}(T)} \gamma_{(\theta, q_{-i})}(\Theta \times q_{-i}^{-1}(B_{\frac{1}{n}}(q_{-i}(t_{-i}))) \times \hat{B}_{\frac{1}{n}}(m_{-i}))d\pi_i^*(q_i(t_i)) \\
&= \gamma(\Theta \times q_{-i}^{-1}(B_{\frac{1}{n}}(q_{-i}(t_{-i}))) \times \hat{B}_{\frac{1}{n}}(m_{-i})).
\end{aligned}$$

So there must exist $t_{-i}^n \in q_{-i}^{-1}(B_{\frac{1}{n}}(q_{-i}(t_{-i})))$ and $m_{-i}^n \in \hat{B}_{\frac{1}{n}}(m_{-i})$ such that (t_{-i}^n, m_{-i}^n) is in the support of the marginal distribution of γ on $T_{-i} \times M_{-i}$, which in turn is a subset of $\times_{j \neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$. Thus, $(t_{-i}^n, m_{-i}^n) \in \times_{j \neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$ for all n. By construction, $m_{-i}^n \to m_{-i}$. As T_{-i} is compact, there exists a subsequence, without loss of generality the sequence $(t_{-i}^n)_{n \geq 1}$ itself, that converges to some $\tilde{t}_{-i}$. Since $\alpha_i^{\mu,\mathcal{T}}$ is continuous^p, it is continuous for all i. Therefore, we can apply Lemma 3 to obtain that $Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$ is closed for all $j \neq i$. Hence, $(\tilde{t}_{-i}, m_{-i}) \in \times_{j \neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$. So $m_{-i} \in \times_{j \neq i} \Sigma_j^{k-1}(\tilde{t}_j|\alpha^{\mu,\mathcal{T}})$.

By construction, $t_{-i}^n \xrightarrow{p} t_{-i}$. Hence, $q_{-i}(\tilde{t}_{-i}) = q_{-i}(t_{-i})$. By the induction hypothesis, $\times_{j \neq i} \Sigma_j^{k-1}(\tilde{t}_j|\alpha^{\mu,\mathcal{T}}) = \times_{j \neq i} \Sigma_j^{k-1}(t_j|\alpha^{\mu,\mathcal{T}})$. Hence, $m_{-i} \in \times_{j \neq i} \Sigma_j^{k-1}(t_j|\alpha^{\mu,\mathcal{T}})$, i.e., $(t_{-i}, m_{-i}) \in \times_{j \neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$. Thus, the marginal of γ' on $T_{-i} \times M_{-i}$ supports a subset of $\times_{j \neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}}))$.

By construction, for all m'_i,

$$\begin{aligned}
&\int_{\Theta \times T_{-i} \times M_{-i}} U_i(\mu(m'_i, m_{-i}), \theta) d\gamma' \\
&= \int_{\Theta \times T_{-i}} \int_{M_{-i}} U_i(\mu(m'_i, m_{-i}), \theta) d\mathrm{marg}_{M_{-i}}\gamma_{(\theta, q_{-i}(t_{-i}))} d\pi_i(t'_i), \\
&= \int_{\Theta \times Q_{-i}(T)} \int_{M_{-i}} U_i(\mu(m'_i, m_{-i}), \theta) d\mathrm{marg}_{M_{-i}}\gamma_{(\theta, q_{-i})} d\pi_i^*(q_i(t'_i)),
\end{aligned}$$

$$= \int_{\Theta\times T_{-i}\times M_{-i}} U_i(\mu(m_i', m_{-i}), \theta) d\gamma.$$

Therefore,

$$m_i \in \arg\max_{m_i' \in M_i} \int_{\Theta\times T_{-i}\times M_{-i}} U_i(\mu(m_i', m_{-i}), \theta) d\gamma'.$$

Hence, $m_i \in \Sigma_i^k(t_i'|\alpha^{\mu,\mathcal{T}})$. We thus conclude that $\Sigma_i^k(t_i|\alpha^{\mu,\mathcal{T}}) = \Sigma_i^k(t_i'|\alpha^{\mu,\mathcal{T}})$. □

Proof of Theorem 1:

Proof. We prove the sufficiency part of the theorem using a series of steps.

Step 1: Recall the construction of $\hat{\mu}$:

$$\hat{\mu}(q^1) = f(\tau(q^1)), \forall q^1 \in \times_{i\in I} Q_i^1(\hat{T}).$$

Using Lemma 1, we continuously extend $\hat{\mu}$ to $(\Delta\Theta)^I$ to obtain the simple mechanism μ, which basically amounts to applying $\hat{\mu}$ after translating messages $m_i \in \Delta\Theta$ into messages $q_i^1 \in Q_i^1(\hat{T})$ using the translation $q_i^1 \in \omega_i(m_i)$ with probability $\xi_{m_i}(q_i^1)$.

For the rest of the proof, let us fix a type space $\mathcal{T}' \supseteq \hat{T}$ and recall that we are assuming that the behavioral anchors $\alpha^{\mu,\mathcal{T}'}$ in the simple mechanism μ are truthful. An important property of truthful anchors that we use below is that they are continuous^p. We will also apply the technical lemmata presented in Appendix A (Lemmata 2, 3, 4, and 5).

Step 2: As μ is continuous, we can apply Lemma 2 which proves that for each i and σ_{-i}, we can find a measurable (pure) strategy σ_i such that $\sigma_i(t_i)$ is type t_i's best response to σ_{-i}. Hence, using induction on k, we can easily argue that $S_i^k(\alpha^{\mu,\mathcal{T}'}) \neq \emptyset$ for all i and $1 \leq k \leq K$.

Step 3: As both μ and the behavioral anchors $\alpha^{\mu,\mathcal{T}'}$ are continuous, we can apply Lemmata 3 and 4. Thus, firstly, for all individuals i and depths of reasoning $k \geq 1$, the correspondence $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}'})$ is upper hemicontinuous. Secondly, for all individuals i, depths of reasoning $k \geq 1$, strategies $\sigma_i \in S_i^k(\alpha^{\mu,\mathcal{T}'})$, and types $t_i \in T_i'$, every m_i in the support of $\sigma_i(t_i)$ is an element of $\Sigma_i^k(t_i|\alpha^{\mu,\mathcal{T}'})$.

Step 4: As the mechanism μ is continuous and the behavioral anchors $\alpha^{\mu,\mathcal{T}'}$ are continuous^p, we can apply Lemma 5. Thus, the correspondence $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}'})$ depends only on the belief hierarchies of types. This result, together with upper hemicontinuity of $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}})$, implies that $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}})$ is in fact upper hemicontinuous^p. We prove this by arguing that $Gr(\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}}))$ is closed when the underlying topology on T_i' is the semimetric topology. Consider any $(t_i^n)_{n\geq 1} \in T_i'$ and $(m_i^n)_{n\geq 1} \in M_i$ such that $t_i^n \xrightarrow{p} t_i$, $m_i^n \to m_i$ and $m_i^n \in \Sigma_i^k(t_i^n|\alpha^{\mu,\mathcal{T}})$ for all $n \geq 1$. As T_i' is compact, there exists a subsequence, without loss of generality the sequence $(t_i^n)_{n\geq 1}$ itself, that converges to some t_i'. Then upper hemicontinuity of $\Sigma_i^k(\cdot|\alpha^{\mu,\mathcal{T}})$ implies that $m_i \in \Sigma_i^k(t_i'|\alpha^{\mu,\mathcal{T}})$. But $t_i^n \to t_i'$ implies that $t_i^n \xrightarrow{p} t_i'$. Thus, we must have $q_i(t_i) = q_i(t_i')$. Lemma 5 then implies that $m_i \in \Sigma_i^k(t_i|\alpha^{\mu,\mathcal{T}})$.

Step 5: We need one more lemma before making the final argument. The following lemma says that for all individuals i, depths of reasoning $k \geq 1$, and types t_i who belong to the planner's model, the messages in $\Sigma_i^k(t_i|\alpha^{\mu,\mathcal{T}'})$ translate into messages in $\{q_i^1(t_i') : t_i' \sim_i^f t_i\}$.

Lemma 6. *For all $\mathcal{T}' \supseteq \hat{T}$, $i \in I$, $t_i \in \hat{T}_i$, and $k \geq 1$, if $m_i \in \Sigma_i^k(t_i|\alpha^{\mu,\mathcal{T}'})$, then $\omega_i(m_i) \subseteq \{q_i^1(t_i') : t_i' \sim_i^f t_i\}$.*

Proof. We first make the following observation: As f is SIRBIC, and $\hat{\mu}(q^1(t)) = f(\tau(q^1(t))) = f(t)$ for all $t \in \hat{T}$, we have

$$\begin{aligned}\int_{\Theta\times\hat{T}_{-i}} U_i(\hat{\mu}(q^1(t)),\theta)d\hat{\pi}_i(t_i) &= \int_{\Theta\times\hat{T}_{-i}} U_i(f(t),\theta)d\hat{\pi}_i(t_i)\\ &\geq \int_{\Theta\times\hat{T}_{-i}} U_i(f(t_i',t_{-i}),\theta)d\hat{\pi}_i(t_i)\\ &= \int_{\Theta\times\hat{T}_{-i}} U_i(\hat{\mu}(q_i^1(t_i'),q_{-i}^1(t_{-i})),\theta)d\hat{\pi}_i(t_i),\end{aligned}$$

for all $t_i, t_i' \in \hat{T}_i$ and $i \in I$, and the inequality holds strictly when f is responsive to t_i versus t_i'.

We now proceed by induction on k. Pick any individual i and type $t_i \in \hat{T}_i$. Since $\alpha^{\mu,\mathcal{T}'}$ is truthful, if $m_i \in \Sigma_i^1(t_i|\alpha^{\mu,\mathcal{T}'}) = BR_i^{\alpha_{-i}^{\mu,\mathcal{T}'}}(t_i)$, then we must have

$$m_i \in \arg\max_{m_i' \in \Delta\Theta} \sum_{q_i^1 \in \omega_i(m_i')} \xi_{m_i'}(q_i^1) \int_{\Theta\times\hat{T}_{-i}} U_i(\hat{\mu}(q_i^1, q_{-i}^1(t_{-i})),\theta)d\hat{\pi}_i(t_i). \tag{6}$$

It follows from the observation above that if $q_i^1 \in \omega_i(m_i)$, then $q_i^1 = q_i^1(t_i')$ for some $t_i' \sim_i^f t_i$. Thus, $\omega_i(m_i) \subseteq \{q_i^1(t_i') : t_i' \sim_i^f t_i\}$.

Suppose now that $k > 1$, and that the property holds for all $k' < k$. Consider any $t_i \in \hat{T}_i$ and $m_i \in \Sigma_i^k(t_i|\alpha^{\mu,\mathcal{T}'})$. Then m_i is a best response to some conjecture $\gamma \in \Delta(\Theta \times T'_{-i} \times M_{-i})$ such that the marginal of γ on $\Theta \times T'_{-i}$ is equal to $\pi_i'(t_i)$ (which is equal to $\hat{\pi}_i(t_i)$ on its support) and the marginal distribution of γ on $T'_{-i} \times M_{-i}$ supports a subset of $\times_{j \neq i} Gr(\Sigma_j^{k-1}(\cdot|\alpha^{\mu,\mathcal{T}'}))$. By the induction hypothesis, individual i's conjecture has j of any type $t_j \in \hat{T}_j$ report m_j such that $\omega_j(m_j) \subseteq \{q_j^1(t_j') : t_j' \sim_j^f t_j\}$. But for any $q_i^1 \in Q_i^1(\hat{T})$ and $t_{-i} \in \hat{T}_{-i}$, the following is true for all t'_{-i} such that $t_j' \sim_j^f t_j$ for all j:

$$\hat{\mu}(q_i^1, q_{-i}^1(t'_{-i})) = f(\tau_i(q_i^1), t'_{-i}) = f(\tau_i(q_i^1), t_{-i}) = \hat{\mu}(q_i^1, q_{-i}^1(t_{-i})).$$

It is thus without loss of generality to assume that individual i's conjecture has j of any type $t_j \in \hat{T}_j$ report m_j such that m_j translates into $q_j^1(t_j)$. So the message m_i must also satisfy (6). Hence, for the same reason as above, we have $\omega_i(m_i) \subseteq \{q_i^1(t_i') : t_i' \sim_i^f t_i\}$, as desired. □

Step 6: To finish the proof, let $(t^n)_{n\geq 1}$ be a sequence of type profiles in T' such that $t^n \xrightarrow{p} t \in \hat{T}$. Pick any strategy profile σ such that for each i, $\sigma_i \in S_i^{k_i}(\alpha^{\mu,\mathcal{T}'})$ with $1 \leq k_i \leq K$.

Consider individual i. As $M_i = \Delta\Theta$ is compact, ΔM_i is compact. Compactness of ΔM_i implies that every *subsequence* of $\sigma_i(t_i^n)$ has a subsequence $\sigma_i(t_i^{n_l})$ that converges to some $\eta_i \in \Delta M_i$. Pick any message m_i in the support of η_i. Since $\sigma_i(t_i^{n_l})$ converges to η_i, and the support correspondence is lower hemicontinuous (see Theorem 17.14 in Aliprantis and Border (2006)), we can find a subsequence of $\sigma_i(t_i^{n_l})$, without loss of generality the sequence $\sigma_i(t_i^{n_l})$ itself, and corresponding sequence of messages $m_i^{n_l}$ such that $m_i^{n_l}$ is in the support of $\sigma_i(t_i^{n_l})$ for each n_l and $m_i^{n_l}$ converges to m_i. Now, $m_i^{n_l} \in \Sigma_i^{k_i}(t_i^{n_l}|\alpha^{\mu,\mathcal{T}'})$, as argued in Step 3. Since $t_i^{n_l} \xrightarrow{p} t_i$, $m_i^{n_l} \to m_i$, and $\Sigma_i^{k_i}(\cdot|\alpha^{\mu,\mathcal{T}'})$ is upper hemicontinuous^p, we have that $m_i \in \Sigma_i^{k_i}(t_i|\alpha^{\mu,\mathcal{T}'})$. It then follows from Lemma 6 that $\omega_i(m_i) \subseteq \{q_i^1(t_i') : t_i' \sim_i^f t_i\}$.

Since we have a finite number of individuals, the previous argument implies that every *subsequence* of $\sigma(t^n)$ has a subsequence $\sigma(t^{n_l})$ that converges to some

$(\eta_1, \dots, \eta_I) \in \times_{i\in I}\Delta M_i$ such that if the message profile m is such that m_i is in the support of η_i for each i, then $\omega_i(m_i) \subseteq \{q_i^1(t_i') : t_i' \sim_i^f t_i\}$ for each i. We argue that $\mu(\sigma(t^{n_l}))$ converges $f(t)$. To see this, pick any Borel subset B of X such that $f(t)(\partial B) = 0$, where ∂B denotes the boundary of B. Then the mapping $m \to \mu(m)(B)$ is continuous (due to the continuity of μ) and bounded. Since $\sigma(t^{n_l})$ converges to $(\eta_1, \dots, \eta_I)$, it follows from the definition of weak converge of probability measures that $\mu(\sigma(t^{n_l}))(B) = \int_M \mu(m)(B)d\sigma_1(t_1^{n_l}) \times \dots \times \sigma_I(t_I^{n_l})$ converges to $\int_M \mu(m)(B)d\eta_1 \times \dots \times \eta_I$. But any m in the support of $\eta_1 \times \dots \times \eta_I$ translates into profiles in $\times_{i\in I}\{q_i^1(t_i') : t_i' \sim_i^f t_i\}$. Since $t' \sim^f t$ implies that $\hat\mu(q_1^1(t_1'), \dots, q_I^1(t_I')) = f(t') = f(t)$, we have that $\mu(m) = f(t)$ for all m in the support of $\eta_1 \times \dots \times \eta_I$. Thus, $\int_M \mu(m)(B)d\eta_1 \times \dots \times \eta_I = f(t)(B)$, and hence $\mu(\sigma(t^{n_l}))(B)$ converges to $f(t)(B)$. Therefore, $\mu(\sigma(t^{n_l}))$ converges to $f(t)$ by the definition of weak convergence of probability measures.

It follows from the argument in the previous paragraph that every *subsequence* of $\mu\circ\sigma(t^n)$ has a subsequence that converges to $f(t)$, which is sufficient to conclude that the sequence $\mu \circ \sigma(t^n)$ itself converges to $f(t)$. □

Proof of Theorem 3:

Proof. Consider an unacceptable deception β. Then $f(\beta(\hat t)) \neq \{f(\hat t)\}$ for some $\hat t \in \hat T$. It follows from Lemma 7 (which is stated and proved at the end of the current proof) that there exist $i \in I$, $t_i \in \hat T_i$, and $\tilde t_i \in \beta_i(t_i)$ such that for every conjecture $\gamma \in \Delta(\Theta \times \hat T_{-i} \times \hat T_{-i}^\beta)$ for which the marginal distribution of γ on $\Theta \times \hat T_{-i}$ coincides with $\hat\pi_i(t_i)$ and $\gamma(\theta, t_{-i}, \tilde t_{-i}) > 0 \Longrightarrow \tilde t_{-i} \in \beta_{-i}(t_{-i})$, there exists $\mathcal{T} \supseteq \hat T$ and $\bar t_i \in T_i$ such that

$$\sum_{\Theta\times\hat T_{-i}^\beta} U_i(f^{\mathcal T}(\bar t_i, t_{-i}), \theta)\ \mathrm{marg}_{\Theta\times\hat T_{-i}^\beta}\gamma(\theta, t_{-i}) > \sum_{\Theta\times\hat T_{-i}^\beta} U_i(f^{\mathcal T}(\tilde t_i, t_{-i}), \theta)\ \mathrm{marg}_{\Theta\times\hat T_{-i}^\beta}\gamma(\theta, t_{-i}).$$

Thus it must be that $f^{\mathcal T}(\bar t_i, t_{-i}) \neq f^{\mathcal T}(\tilde t_i, t_{-i})$ for at least one $t_{-i} \in \hat T_{-i}^\beta$. Therefore, $\tilde t_i \not\sim_i^{f^{\mathcal T}} \bar t_i$.

Define $\ell : \hat T_{-i} \to \Delta X$ as $\ell(t_{-i}) = f^{\mathcal T}(\bar t_i, t_{-i})$ for all $t_{-i} \in \hat T_{-i}$. Since $\tilde t_i \in \hat T_i$, we have $f^{\mathcal T}(\tilde t_i, t_{-i}) = f(\tilde t_i, t_{-i})$ for all $t_{-i} \in \hat T_{-i}^\beta$. Therefore, the above inequality can be rewritten as

$$\sum_{\Theta\times\hat{T}^{\beta}_{-i}} U_i(\ell(t_{-i}),\theta)\ \mathrm{marg}_{\Theta\times\hat{T}^{\beta}_{-i}}\gamma(\theta,t_{-i}) > \sum_{\Theta\times\hat{T}^{\beta}_{-i}} U_i(f(\tilde{t}_i,t_{-i}),\theta)\ \mathrm{marg}_{\Theta\times\hat{T}^{\beta}_{-i}}\gamma(\theta,t_{-i}).$$

As $f^{\mathcal{T}}$ is Bayesian incentive compatible, for all $i\in I$ and $t_i'\in\hat{T}_i$,

$$\begin{aligned}\sum_{\Theta\times\hat{T}_{-i}} U_i(f(t_i',t_{-i}),\theta)\hat{\pi}_i(t_i')(\theta,t_{-i}) &= \sum_{\Theta\times\hat{T}_{-i}} U_i(f^{\mathcal{T}}(t_i',t_{-i}),\theta)\hat{\pi}_i(t_i')(\theta,t_{-i})\\ &\geq \sum_{\Theta\times\hat{T}_{-i}} U_i(f^{\mathcal{T}}(\bar{t}_i,t_{-i}),\theta)\hat{\pi}_i(t_i')(\theta,t_{-i})\\ &= \sum_{\Theta\times\hat{T}_{-i}} U_i(\ell(t_{-i}),\theta)\hat{\pi}_i(t_i')(\theta,t_{-i}).\end{aligned}$$

Finally, since $f^{\mathcal{T}}$ is SIRBIC with respect to types in $\hat{T}$ and $\tilde{t}_i \not\sim_i^{f^{\mathcal{T}}} \bar{t}_i$, the above inequality must be strict when $t_i'=\tilde{t}_i$. □

We finish the appendix with the next lemma and its proof:

Lemma 7. *Suppose $f:\hat{T}\to\Delta X$ can be extended to every type space $\mathcal{T}\supseteq\hat{T}$ such that the extension $f^{\mathcal{T}}$ is Bayesian incentive compatible, SIRBIC with respect to types in $\hat{T}$, and continuous^p at all points in $\hat{T}$. If there exists a deception β such that for each $i\in I$, $t_i\in\hat{T}_i$, and $\tilde{t}_i\in\beta_i(t_i)$, there exists a conjecture $\gamma^{\tilde{t}_i}_{t_i}\in\Delta(\Theta\times\hat{T}_{-i}\times\hat{T}^{\beta}_{-i})$ such that the marginal distribution of $\gamma^{\tilde{t}_i}_{t_i}$ on $\Theta\times\hat{T}_{-i}$ coincides with $\hat{\pi}_i(t_i)$, $\gamma^{\tilde{t}_i}_{t_i}(\theta,t_{-i},\tilde{t}_{-i})>0\implies\tilde{t}_{-i}\in\beta_{-i}(t_{-i})$ and*

$$\tilde{t}_i\in\arg\max_{t_i'\in T_i}\sum_{\Theta\times\hat{T}^{\beta}_{-i}} U_i(f^{\mathcal{T}}(t_i',t_{-i}),\theta)\ \mathrm{marg}_{\Theta\times\hat{T}^{\beta}_{-i}}\gamma^{\tilde{t}_i}_{t_i}(\theta,t_{-i}),$$

for all $\mathcal{T}\supseteq\hat{T}$, then $f(\beta(\hat{t}))=\{f(\hat{t})\}$ for all $\hat{t}\in\hat{T}$.

Proof. Consider any $t_i\in\hat{T}_i$ and $\tilde{t}_i\in\beta_i(t_i)$. Since $\tilde{t}_i\in\hat{T}_i$, and f is SIRBIC, the set of best replies of type $\tilde{t}_i$ to the belief that all other individuals report their types truthfully in the direct mechanism f is equal to $\{t_i'\in\hat{T}_i : t_i'\sim_i^f\tilde{t}_i\}$. Let $\gamma^{\tilde{t}_i}_{*}\in\Delta(\Theta\times\hat{T}_{-i}\times\hat{T}_{-i})$ denote the conjecture corresponding to this belief of type $\tilde{t}_i$. Thus, $\gamma^{\tilde{t}_i}_{*}(\theta,t_{-i},t_{-i})=\hat{\pi}_i(\tilde{t}_i)(\theta,t_{-i})$ and $\gamma^{\tilde{t}_i}_{*}(\theta,t_{-i},t_{-i}')=0$ for all $t_{-i}'\neq t_{-i}$.

For each $z \geq 1$, $i \in I$, $t_i \in \hat{T}_i$ and $\tilde{t}_i \in \beta_i(t_i)$, we construct $\bar{t}_i[z, t_i, \tilde{t}_i]$ and let $\mathcal{T}$ be the type space such that

$$T_i = \bigcup_{z \geq 1} \bigcup_{t_i \in \hat{T}_i} \bigcup_{\tilde{t}_i \in \beta_i(t_i)} \{\bar{t}_i[z, t_i, \tilde{t}_i]\} \bigcup \hat{T}_i, \forall i \in I.$$

For each i, pick any $|\hat{T}_i|$ points in $\Re^2$ such that the Euclidean distance between any two points is exactly 1. Since $\hat{T}_i$ is homeomorphic to this set of points, we can identify each point in this set by the corresponding $t_i \in \hat{T}_i$. For each $t_i \in \hat{T}_i$, draw a circle of radius $\frac{1}{4}$ in $\Re^2$ with the center at t_i. For each $\tilde{t}_i \in \beta_i(t_i)$, pick a distinct point on the circumference of this circle, and denote this point by $\bar{t}_i[1, t_i, \tilde{t}_i]$ (note that there are a finite number of such points). For each $z > 1$, let $\bar{t}_i[z, t_i, \tilde{t}_i]$ be the point on the line-segment joining t_i and $\bar{t}_i[1, t_i, \tilde{t}_i]$ which is at a distance $\frac{1}{3+z}$ from t_i. By construction, T_i is countable, and when endowed with the Euclidean metric, it is a compact metric space.

For each $t_i \in \hat{T}_i$, we let the belief $\pi_i(t_i) = \hat{\pi}_i(t_i)$. Then define the beliefs of type $\bar{t}_i[1, t_i, \tilde{t}_i]$ such that $\pi_i(\bar{t}_i[1, t_i, \tilde{t}_i]) = \pi_i(\tilde{t}_i)$. For each $z > 1$, the beliefs of type $\bar{t}_i[z, t_i, \tilde{t}_i]$ are such that

$$\pi_i(\bar{t}_i[z, t_i, \tilde{t}_i])(\theta, \bar{t}_{-i}[z-1, t_{-i}, \tilde{t}_{-i}]) = \left(1 - \frac{1}{z}\right) \gamma_{t_i}^{\tilde{t}_i}(\theta, t_{-i}, \tilde{t}_{-i}) + \frac{1}{z} \gamma_*^{\tilde{t}_i}(\theta, t_{-i}, \tilde{t}_{-i}),$$

for all $t_{-i} \in \hat{T}_{-i}$ and $\tilde{t}_{-i} \in \beta_{-i}(t_{-i})$, and

$$\pi_i(\bar{t}_i[z, t_i, \tilde{t}_i])(\theta, t_{-i}) = \frac{1}{z} \gamma_*^{\tilde{t}_i}(\theta, t_{-i}, t_{-i}),$$

for all $t_{-i} \in \hat{T}_{-i}$ such that $t_{-i} \notin \beta_{-i}(t_{-i})$.

We argue that π_i is continuous. As all points in T_i except those in $\hat{T}_i$ are isolated, we only need to show that π_i is continuous at all $t_i \in \hat{T}_i$. It is sufficient to argue that $\pi_i(\bar{t}_i[z, t_i, \tilde{t}_i])$ converges to $\pi_i(t_i)$ as $z \to \infty$ for all $\tilde{t}_i \in \beta_i(t_i)$. To see this, consider any bounded continuous function $h : \Theta \times T_{-i} \to \Re$. Then for any $z > 1$,

$$\sum_{\Theta \times T_{-i}} h(\theta, t_{-i}) \pi_i(\bar{t}_i[z, t_i, \tilde{t}_i])(\theta, t_{-i})$$

$$= \sum_{(\theta,t_{-i},\tilde{t}_{-i}):\tilde{t}_{-i}\in\beta_{-i}(t_{-i})} h(\theta,\bar{t}_{-i}[z-1,t_{-i},\tilde{t}_{-i}])\left(\left(1-\frac{1}{z}\right)\gamma_{t_i}^{\tilde{t}_i}(\theta,t_{-i},\tilde{t}_{-i})+\frac{1}{z}\gamma_*^{\tilde{t}_i}(\theta,t_{-i},\tilde{t}_{-i})\right)$$
$$+ \sum_{(\theta,t_{-i}):t_{-i}\notin\beta_{-i}(t_{-i})} h(\theta,t_{-i})\frac{1}{z}\gamma_*^{m_i}(\theta,t_{-i},t_{-i}).$$

As h is continuous and $\lim_{z\to\infty}\bar{t}_{-i}[z-1,t_{-i},\tilde{t}_{-i}]=t_{-i}$, we have

$$\begin{aligned}\lim_{z\to\infty}\sum_{\Theta\times T_{-i}} h(\theta,t_{-i})\pi_i(\bar{t}_i[z,t_i,\tilde{t}_i])(\theta,t_{-i}) &= \sum_{(\theta,t_{-i},\tilde{t}_{-i}):\tilde{t}_{-i}\in\beta_{-i}(t_{-i})} h(\theta,t_{-i})\gamma_{t_i}^{\tilde{t}_i}(\theta,t_{-i},\tilde{t}_{-i})\\ &= \sum_{(\theta,t_{-i})} h(\theta,t_{-i})\pi_i(t_i)(\theta,t_{-i})\end{aligned}$$

Thus, $\pi_i(\bar{t}_i[z,t_i,\tilde{t}_i])$ converges to $\pi_i(t_i)$ as $z\to\infty$.

By construction, $\mathcal{T}\supseteq\hat{T}$. By assumption, f can be extended to $\mathcal{T}$ such that the extension $f^{\mathcal{T}}$ is Bayesian incentive compatible, SIRBIC with respect to types in $\hat{T}$, and continuous^p at all points in $\hat{T}$. We argue that for all $z\geq 1$, $t_i\in\hat{T}_i$ and $\tilde{t}_i\in\beta_i(t_i)$,

$$f^{\mathcal{T}}(\bar{t}_i[z,t_i,\tilde{t}_i]),t_{-i})=f^{\mathcal{T}}(\tilde{t}_i,t_{-i}),\forall t_{-i}\in T_{-i}.$$

We proceed by induction. Consider $\bar{t}_i[1,t_i,\tilde{t}_i]$. Since $f^{\mathcal{T}}$ is Bayesian incentive compatible, $\bar{t}_i[1,t_i,\tilde{t}_i]$ is a best response of type $\bar{t}_i[1,t_i,\tilde{t}_i]$ when all others report their types truthfully in the direct mechanism $f^{\mathcal{T}}$. By construction, $\pi_i(\bar{t}_i[1,t_i,\tilde{t}_i])=\pi_i(\tilde{t}_i)$. Thus, the set of best responses of type $\bar{t}_i[1,t_i,\tilde{t}_i]$ are equal to the set of best responses of type $\tilde{t}_i$ when all others report their types truthfully in the direct mechanism $f^{\mathcal{T}}$. This means that $\bar{t}_i[1,t_i,\tilde{t}_i]$ too is a best response of type $\tilde{t}_i$ when all others report their types truthfully in the direct mechanism $f^{\mathcal{T}}$. Since $f^{\mathcal{T}}$ is SIRBIC with respect to types in $\hat{T}$, we obtain that

$$f^{\mathcal{T}}(\bar{t}_i[1,t_i,\tilde{t}_i]),t_{-i})=f^{\mathcal{T}}(\tilde{t}_i,t_{-i}),\forall t_{-i}\in T_{-i}.$$

Now pick $z\geq 1$, and suppose the statement is true for $z-1$. Consider $\bar{t}_i[z,t_i,\tilde{t}_i]$. If $\bar{t}_i[z,t_i,\tilde{t}_i]$ reports $t'_i\in T_i$ when all others report their type truthfully in the direct mechanism $f^{\mathcal{T}}$, then he expects a payoff of

$$\begin{aligned}
&\sum_{\Theta\times T_{-i}} U_i(f^{\mathcal{T}}(t_i', t_{-i}), \theta)\pi_i(\bar{t}_i[z, t_i, \tilde{t}_i])(\theta, t_{-i}) \\
&= \sum_{(\theta, t_{-i}, \tilde{t}_{-i}):\tilde{t}_{-i}\in\beta_{-i}(t_{-i})} U_i(f^{\mathcal{T}}(t_i', \bar{t}_{-i}[z-1, t_{-i}, \tilde{t}_{-i}]), \theta)\left(\left(1-\frac{1}{z}\right)\gamma_{t_i}^{\tilde{t}_i}(\theta, t_{-i}, \tilde{t}_{-i}) + \frac{1}{z}\gamma_*^{\tilde{t}_i}(\theta, t_{-i}, \tilde{t}_{-i})\right) \\
&\quad + \sum_{(\theta, t_{-i}):t_{-i}\notin\beta_{-i}(t_{-i})} U_i(f^{\mathcal{T}}(t_i', t_{-i}), \theta)\frac{1}{z}\gamma_*^{\tilde{t}_i}(\theta, t_{-i}, t_{-i}). \qquad (7)
\end{aligned}$$

By the induction hypothesis, $f^{\mathcal{T}}(t_i', \bar{t}_{-i}[z-1, t_{-i}, \tilde{t}_{-i}]) = f^{\mathcal{T}}(t_i', \tilde{t}_{-i})$. Hence, the right-hand side of (7) becomes

$$\begin{aligned}
&\sum_{(\theta, t_{-i}, \tilde{t}_{-i}):\tilde{t}_{-i}\in\beta_{-i}(t_{-i})} U_i(f^{\mathcal{T}}(t_i', \tilde{t}_{-i}), \theta)\left(\left(1-\frac{1}{z}\right)\gamma_{t_i}^{\tilde{t}_i}(\theta, t_{-i}, \tilde{t}_{-i}) + \frac{1}{z}\gamma_*^{\tilde{t}_i}(\theta, t_{-i}, \tilde{t}_{-i})\right) \\
&\quad + \sum_{(\theta, t_{-i}):t_{-i}\notin\beta_{-i}(t_{-i})} U_i(f^{\mathcal{T}}(t_i', t_{-i}), \theta)\frac{1}{z}\gamma_*^{\tilde{t}_i}(\theta, t_{-i}, t_{-i}) \\
&= \left(1-\frac{1}{z}\right)\sum_{\Theta\times\hat{T}_{-i}^{\beta}} U_i(f^{\mathcal{T}}(t_i', \tilde{t}_{-i}), \theta)\ \mathrm{marg}_{\Theta\times\hat{T}_{-i}^{\beta}}\gamma_{t_i}^{\tilde{t}_i}(\theta, \tilde{t}_{-i}) + \frac{1}{z}\sum_{\Theta\times\hat{T}_{-i}} U_i(f^{\mathcal{T}}(t_i', t_{-i}), \theta)\pi_i(\tilde{t}_i)(\theta, t_{-i})
\end{aligned}$$

By assumption,

$$\tilde{t}_i \in \arg\max_{t_i'\in T_i}\sum_{\Theta\times\hat{T}_{-i}^{\beta}} U_i(f^{\mathcal{T}}(t_i', \tilde{t}_{-i}), \theta)\ \mathrm{marg}_{\Theta\times\hat{T}_{-i}^{\beta}}\gamma_{t_i}^{\tilde{t}_i}(\theta, \tilde{t}_{-i}).$$

Since $f^{\mathcal{T}}$ is Bayesian incentive compatible,

$$\tilde{t}_i \in \arg\max_{t_i'\in T_i}\sum_{\Theta\times\hat{T}_{-i}} U_i(f^{\mathcal{T}}(t_i', t_{-i}), \theta)\pi_i(\tilde{t}_i)(\theta, t_{-i}).$$

Then it must be that

$$\bar{t}_i[z, t_i, \tilde{t}_i] \in \arg\max_{t_i'\in T_i}\sum_{\Theta\times\hat{T}_{-i}} U_i(f^{\mathcal{T}}(t_i', t_{-i}), \theta)\pi_i(\tilde{t}_i)(\theta, t_{-i}). \qquad (8)$$

If that were not true, then

$$\sum_{\Theta\times T_{-i}} U_i(f^{\mathcal{T}}(\tilde{t}_i, t_{-i}), \theta)\pi_i(\bar{t}_i[z, t_i, \tilde{t}_i])(\theta, t_{-i}) > \sum_{\Theta\times T_{-i}} U_i(f^{\mathcal{T}}(\bar{t}_i[z, t_i, \tilde{t}_i], t_{-i}), \theta)\pi_i(\bar{t}_i[z, t_i, \tilde{t}_i])(\theta, t_{-i}),$$

which would contradict the fact that $f^{\mathcal{T}}$ is Bayesian incentive compatible.

The expression (8) means that $\bar{t}_i[z, t_i, \tilde{t}_i]$ too is a best response of type $\tilde{t}_i$ when all others report their types truthfully in the direct mechanism $f^{\mathcal{T}}$. Since $f^{\mathcal{T}}$ is SIRBIC with respect to types in $\hat{T}$, we obtain that

$$f^{\mathcal{T}}(\bar{t}_i[z, t_i, \tilde{t}_i]), t_{-i}) = f^{\mathcal{T}}(\tilde{t}_i, t_{-i}), \forall t_{-i} \in T_{-i}.$$

To complete the proof, consider any $t \in \hat{T}$ and $\tilde{t} \in \beta(t)$. We have argued that

$$f^{\mathcal{T}}(\bar{t}[z, t, \tilde{t}]) = f^{\mathcal{T}}(\tilde{t}), \forall z \geq 1.$$

Since $\bar{t}[z, t, \tilde{t}] \to t$, we have $\bar{t}[z, t, \tilde{t}] \xrightarrow{p} t$. As $f^{\mathcal{T}}$ is continuous^p at all points in $\hat{T}$, we have $f^{\mathcal{T}}(\tilde{t}) = f^{\mathcal{T}}(t)$. But $f^{\mathcal{T}}(\tilde{t}) = f(\tilde{t})$ and $f^{\mathcal{T}}(t) = f(t)$ since $f^{\mathcal{T}}$ is an extension of f. Hence $f(\tilde{t}) = f(t)$. □

References

Aghion, P., D. Fudenberg, R. Holden, T. Kunimoto, and O. Tercieux (2012), "Subgame-perfect Implementation under Information Perturbations," *Quarterly Journal of Economics* 127, 1843-1881.

Aliprantis, C. D. and K. C. Border (2006), "Infinite Dimensional Analysis: A Hitchhiker's Guide," *Springer-Verlag.*

Arens, R. (1952), "Extension of Functions on Fully Normal Spaces," *Pacific Journal of Mathematics* 2, 11-22.

Artemov, G., T. Kunimoto, and R. Serrano (2013), "Robust Virtual Implementation: Toward a Reinterpretation of the Wilson Doctrine," *Journal of Economic Theory* 148, 424-447.

Berge, C. (1963), *Topological Spaces*, New York: Macmillan.

Bergemann, D. and S. Morris (2005), "Robust Mechanism Design," *Econometrica* 73, 1771-1813.

Bergemann, D. and S. Morris (2012), *Robust Mechanism Design,* World Scientific Publishing, Singapore.

Binmore, K., J. McCarthy, G. Ponti, L. Samuelson, and A. Shaked (2002), "A Backward Induction Experiment," *Journal of Economic Theory* 104, 48-88.

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

Brandenburger, A., and E. Dekel (1993). "Hierarchies of Beliefs and Common Knowledge," *Journal of Economic Theory* 59, 189-198.

Chung, K. and J. Ely (2003), "Implementation with Near-Complete Information," *Econometrica* 71, 857-871.

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. (2021). "Efficient Mechanisms for Level-k Bilateral Trading," *Games and Economic Behavior*, `https://doi.org/10.1016/j.geb.2021.02.005` (forthcoming).

d'Aspremont, C. and L.-A. Gerard-Varet (1979), "Incentives and Incomplete Information," *Journal of Public Economics* 11, 25-45.

de Clippel, G., R. Saran, and R. Serrano (2014), "Mechanism Design with Bounded Depth of Reasoning and Small Modeling Mistakes," Unpublished Working Paper 2014-7, Department of Economics, Brown University.

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

Dekel, E., D. Fudenberg, and S. Morris (2007), "Interim Correlated Rationalizability," *Theoretical Economics* 2, 15-40.

Di Tillio, A. (2011), "A Robustness Result for Rationalizable Implementation," *Games and Economic Behavior* 72, 301-305.

Dugundji, J. (1951), "An Extension of Tietze's Theorem," *Pacific Journal of Mathematics* 1, 353-367.

Harsanyi, J. C. (1967, 1968), "Games of Incomplete Information Played by Bayesian Players (parts I, II, and III)," *Management Science* 14, 159-182, 320-334, and 486-

502.

Heifetz, A. and W. Kets (2018). “Robust Multiplicity with a Grain of Naiveté,” *Theoretical Economics* 13, 415-465.

Heifetz, A. and Z. Neeman (2006). “On the Generic (Im)Possibility of Full Surplus Extraction in Mechanism Design,” *Econometrica* 74, 213-233.

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.

Jehiel, P., M. Meyer-ter-Vehn, and B. Moldovanu (2012), “Locally Robust Implementation and its Limits,” *Journal of Economic Theory* 147, 2439-2452.

Katok, E., M. Sefton, and A. Yavas (2002). “Implementation by Iterative Dominance and Backward Induction: An Experimental Comparison,” *Journal of Economic Theory* 104, 89-103.

Kets, W. (2017). “Bounded Reasoning and Higher-Order Uncertainty,” Mimeo, University of Oxford.

Kneeland, T. (2020), “Mechanism Design with Level-k Types: Theory and an Application to Bilateral Trade,” Mimeo, University College London.

Lopomo, G., L. Rigotti, and C. Shannon (2020). “Uncertainty in Mechanism Design,” Mimeo, University of Pittsburgh.

Maskin, E. (1999), “Nash Equilibrium and Welfare Optimality,” *Review of Economic Studies*, vol. 66, 23-38.

McLean, R. P. and A. Postlewaite (2002). “Informational Size and Incentive Compatibility,” *Econometrica* 70, 2421-2453.

Mertens, J.-F. and S. Zamir (1985). “Formulation of Bayesian Analysis for Games of Incomplete Information,” *International Journal of Game Theory* 14, 1-29.

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.

Neeman, Z. (2004), “The Relevance of Private Information in Mechanism Design,” *Journal of Economic Theory* 117, 55-77.

Ollár, M. and A. Penta (2017). “Full Implementation and Belief Restrictions,” *American Economic Review* 107, 2243-2277.

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

Rapoport, A., and W. Amaldoss (2000). "Mixed Strategies and Iterative Elimination of Strongly Dominated Strategies: An Experimental Investigation of States of Knowledge," *Journal of Economic Behavior and Organization* 42, 483-521.

Weinstein, J. and M. Yildiz (2007), "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements," *Econometrica* 75, 365-400.