Mechanism Design with Bounded Depth of Reasoning and Small Modeling Mistakes*

Geoffroy de Clippel† Rene Saran‡ Roberto Serrano§

June 2014

Abstract

We consider mechanism design in contexts in which agents exhibit bounded depth of reasoning (level k) instead of rational expectations. We use simple direct mechanisms, in which agents report only first-order beliefs. While level 0 agents are assumed to be truth tellers, level k agents best-respond to their belief that other agents have at most $k-1$ levels of reasoning. We find that incentive compatibility is necessary for implementation in this framework, while its strict version alone is sufficient. Adding continuity to both directions, the same results are obtained for continuous implementation with respect to small modeling mistakes. We present examples to illustrate the permissiveness of our findings in contrast to earlier related results under the assumption of rational expectations.

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

*We thank Jack Fanning, Laurent Mathevet, Stephen Morris, Arunava Sen, and audience at NYU, ISI (Delhi), and SCW 2014 (Boston) for their comments and suggestions.
†Brown University, Department of Economics, declippel@brown.edu
‡Yale-NUS College, Division of Social Science, rene.saran@yale-nus.edu.sg
§Brown University, Department of Economics, roberto_serrano@brown.edu

1 Introduction

Building institutions that are resilient to misspecifications of basic assumptions is an important task for economists. In the mechanism design literature, concerned with the exploration of institutions in which the informational constraints of the designer are incorporated into the analysis, such resilience or robustness has been addressed in several ways. Bergemann and Morris (2005, 2012) take a global view, by requiring that implementation be robust to any higher-order belief consistent with the underlying payoff-relevant environment. Artemov *et al.* (2013) do not go that far and require robustness to some, but not necessarily all, higher-order beliefs, under the premise that some beliefs can sometimes be discarded due to the planner's prior information. In a context of Knightian uncertainty, Lopomo *et al.* (2009) investigate the limitations that such incomplete preferences may impose on mechanism design. All these approaches model versions of nonlocal robustness, in the sense that the model is tested against a wide class of misspecifications, including some that can potentially be very large.

The current study follows a different approach to robustness, which 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) and Jehiel *et al.* (2013).[1]

But what sets this paper aside from all the work mentioned so far is the change in the agents' behavioral assumptions. In an attempt to endow our theory with further realism, we impose that agents have bounded depth of reasoning. 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.

[1] Other robustness checks 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.

We rely on simple direct mechanisms, in which agents report their first-order beliefs. We assume that our agents perform up to k levels of reasoning, where k can be any nonnegative integer. Level 0 agents are truth-tellers, while level k' agents for any $k' \leq k$ best-respond to their beliefs, which are unrestricted except that they believe that all other agents are of level strictly less than k'. We thus allow for any strategy profile that is consistent with any reasoning of level $k' \leq k$, as just specified, and require that implementation obtain in such strategies.

The bounded depth of reasoning assumption is made with realism in mind. Introspection tells us that long chains of conditional reasoning are hard to perform. The game of chess remains interesting despite being zero-sum (and thus unequivocally solvable from a theoretical perspective) only because of our limited depth of reasoning. Multiple experimental studies in various contexts suggest that people's depth of reasoning is in fact limited.[2]

The way our solution concept is defined may remind the reader of the notion of rationalizability. The main difference between our behavioral model and interim correlated rationalizability (Dekel *et al.* (2007)) is that all our agents' cognitive states start with truth-telling. While it is hard to propose an obvious anchoring point for chains of reasoning in general games, truth-telling seems natural in simple direct mechanisms. Earlier experiments on sender-receiver games provide some support for this intuition (see e.g. Cai and Wang (2006), and Wang *et al.* (2010)). In mechanism design, it is also sometimes argued that truth-telling may constitute a focal point when other equilibria exist. The central role of truth-telling makes our notion of implementation somewhat closer to that of *weak* implementation.

As it turns out, we show that any social choice function (SCF) that is implementable with bounded depth of reasoning must be Bayesian incentive compatible (Theorem 1a). Conversely, any strictly incentive compatible SCF is implementable with bounded depth of reasoning (Theorem 1b). Thus,

[2]See, e.g., Rapaport 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.

even though our behavioral assumption is far removed from equilibrium logic, Bayesian incentive compatibility arises as a robust limitation to the success of mechanism design.

In the second part of our study, we turn to the issue of allowing the planner to make small modeling mistakes. Indeed, the model may not be exactly one of independent types, or private values, or complete information, or selfish agents, to give a few examples of what we might have originally assumed, but an approximation thereof. In this context, we then seek for continuous implementation, much along the lines proposed in Oury and Tercieux (2012), and here is where we find a result that some may deem surprising. Whereas the Oury-Tercieux analysis seems to suggest that the requirement of continuous implementation imposes stringent restrictions (implying a condition that, for instance in complete information environments, is stronger than Maskin monotonicity), our finding here is very different.[3] Other than the requirement of continuity, no additional conditions on top of incentive compatibility are found. Specifically, Theorems 2a and 2b provide an exact counterpart to Theorems 1a and 1b, respectively, by turning the original results of implementation with bounded depth of reasoning into continuous versions of the same form of implementation. It follows that the Oury-Tercieux conclusion hinges on their use of Bayesian equilibrium logic, with its implied unbounded depth of reasoning.

This paper contributes to a growing literature on mechanism design with bounded rationality. These include for instance Eliaz (2002), who studies full implementation in Nash equilibrium that is robust to the presence of any number of "faulty" individuals below a fixed threshold, where faulty individuals may behave in any arbitrary way; Cabrales and Serrano (2011), who investigate implementation problems under the behavioral assumption that

[3]Related to the Oury-Tercieux's logic, see a result for rationalizable implementation of SCFs in Bergemann *et al.* (2011). Indeed, these results teach us, respectively, that continuous weak implementation in strict equilibria or rationalizable implementation of SCFs take us close to full Nash implementation; for a point of comparison, Matsushima (1993) shows that in quasilinear environments full Bayesian implementation comes close to weak implementation, since Bayesian monotonicity is trivially satisfied.

agents myopically adjust their actions in the direction of better-responses or best-responses and derive implementation results for strict equilibria, often found in learning and evolutionary approaches; Saran (2011), who studies under which conditions over individual choice correspondences over Savage acts does the revelation principle hold for partial Nash implementation with incomplete information; Glazer and Rubinstein (2012), who introduce a mechanism design model in which both the content and framing of the mechanism affect the agent's ability to manipulate the information he provides; de Clippel (forthcoming), who studies full Nash implementation when individual choices need not be compatible with preference maximization; and, concurrent to this paper, Saran (2014), who studies implementation under complete information and k levels of rationality. In the present paper, individual behavior is consistent with rationality to the extent that choices emerge from preference maximization given beliefs. However, bounded depth of reasoning relaxes the assumption of rational expectations that underlies the concept of Bayesian Nash equilibrium, which requires the players' beliefs to be consistent with equilibrium behavior. A first investigation of the impact of level k behavior in mechanism design can be found in Crawford *et al.* (2009) who provide some insight on the design of optimal auctions when individuals' depth of reasoning is bounded.

The paper proceeds as follows. Section 2 introduces a motivating example of bilateral trading. Section 3 presents the model. Section 4 presents our central results of implementation with bounded depth of reasoning. Section 5 introduces small modeling mistakes and extends our previous results to continuous implementation. Section 6 showcases three important examples of applications of our results, and Section 7 concludes. The proofs of three key lemmata are relegated to an appendix.

2 A Motivating Example

Consider a simple bilateral trade problem where the seller's good can be of low or high 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. It is thus mutually beneficial to trade the good whatever its quality. Fairness would suggest that the price should fall half-way between the traders' reservation prices. If quality is common knowledge between the buyer and the seller, then a simple direct mechanism makes efficient fair trade compatible with Nash equilibrium. In this mechanism, both the buyer and the seller are asked to report the good's quality, trade occurs if and only if both parties' reports agree, and the good is traded against \$25 (*resp.* \$30) if both parties claim the good is of low (*resp.* high) quality. Clearly, truth-telling is a Nash equilibrium whatever the good's quality, and our desired SCF is (weakly) Nash implementable.

According to Oury and Tercieux (2012), this result is not robust to small modeling mistakes. In particular, it is impossible to find a Bayesian Nash equilibrium of our simple direct mechanism whose resulting outcomes coincide with the desired SCF should the information be complete, and fall close to the desired SCF at nearby information states. To see this, consider types that correspond to infinite hierarchies of deterministic beliefs. If i denotes either the buyer or the seller, then i's type $t_i = (\theta_n)_{n\geq 0}$, where $\theta_n \in \{Low, High\}$ for all n, is interpreted as i believing that the good's quality is θ_0, that $-i$ believes that the good's quality is θ_1, that $-i$ believes that i believes that the good's quality is θ_2, etc. In particular, for each $z \geq 0$, let t^z be the type with $\theta_n^z = High$ for all $n < z$ and $\theta_n^z = Low$ for all $n \geq z$. Notice that t^0 captures either agent's information in the complete information case described in the previous paragraph when the good is of low quality. Consider now a Bayesian Nash equilibrium of our simple direct mechanism. For this equilibrium to implement the desired SCF when information is complete, it must be that either agent reports '*Low*' when his type is t^0. Notice, though, that this implies that an agent of type t^1 expects his opponent to report '*Low*' in that equilibrium, and thus also reports '*Low*'. Iterating this reasoning, we see that both agents must report '*Low*' in that equilibrium when his type is t^z, for all $z \geq 0$. However, t^z converges to the constant type $t^H = (High, High, \dots)$, which captures either agent's information in the complete information case described in the

previous paragraph when the good is of high quality. Hence any Bayesian Nash equilibrium of our simple direct mechanism that delivers the desired outcomes when quality is commonly known must deliver some outcomes that are far from the desired SCF at some nearby information states.[4]

This elegant reasoning rests on the presumption that an agent's behavior can depend on very high-order beliefs, with the agent reporting 'Low' for instance when of type t^{100}, an information state that coincides with knowledge of high quality for one hundred levels of reasoning, while also assuming that this same agent would report '$High$' when the quality is commonly known to be high. While it is most useful to understand as a benchmark what rationality implies when taken to its limit, there is also value in understanding perhaps more realistic circumstances where participants' sophistication is limited. Of course, the proper discussion of such circumstances requires breaking away from the assumption of rational expectations that underlies the concept of Nash equilibrium. The concept of level k play provides a natural place to start in view of the recent literature aiming at better describing the behavior of inexperienced players in games.

To see how the level k model starting with truth-telling at level 0 might help in continuous implementation, reconsider the sequence of types $(t^z)_{z\geq 0}$ converging to t^H in the above example. Level 0 of type t^0 truthfully reports 'Low' as he believes that the good is of low quality whereas for each $z \geq 1$, level 0 of type t^z truthfully reports '$High$' as he believes that the good is of high quality. Type t^0 believes that the other agent is also of type t^0 whereas for each $z > 0$, type t^z believes that the other agent is of type t^{z-1}. Therefore, for each $z \leq 1$, level 1 of type t^z reports 'Low' as he believes that the other agent is of level 0 with type t^0 who truthfully reports 'Low' whereas for each $z \geq 2$, level 1 of type t^z reports '$High$' as he believes that the other agent is of level 0 with type t^{z-1} who truthfully reports '$High$'. By continuing in this fashion, it is easy to argue that for each $z \leq k$, level k of type t^z reports

[4]Oury and Tercieux's result further implies that continuous implementation cannot be achieved in this bilateral trade example, *whatever the mechanism one considers*. They show indeed that continuous implementation requires a form of Maskin monotonicity which is not satisfied by our desired SCF.

'*Low*' whereas for each $z \geq k+1$, level k of type t^z reports '$High$'. Hence if the depth of reasoning is bounded by k, then all types t^z with $z \geq k+1$ at all levels $k' \leq k$ report '$High$', thus preserving continuity with respect to behavior at the limit point t^H.

3 The Model

For any topological space Y, we let ΔY denote the set of probability measures defined on the Borel sigma algebra of subsets of Y. We endow ΔY with the weak* topology, i.e., the topology of weak convergence. If Y is a compact metric space, then ΔY is compact and metrizable by the Prohorov metric.[5] We use the product topology for all product spaces.

3.1 Alternatives, States, and Utility Functions

A social planner/mechanism designer needs to select an alternative from a set X, which we assume to be a compact metric space. His decision impacts the satisfaction of individuals in a finite set I. Unfortunately he 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 that we assume to be also a compact metric space. Individual i evaluates any $l \in \Delta X$ by its expected utility $U_i(l, \theta) = \int_{x \in X} u_i(x, \theta) dl$.

3.2 Information and Beliefs

Let $\mathcal{T} = (T_i^*, \pi_i)_{i \in I}$ be the *universal type space* generated by Θ (see Mertens and Zamir (1985); Brandenburger and Dekel (1993)). Remember that the set T_i^* of individual i's *types* is compact and metrizable, and that the homeomor-

[5] Let d be the metric on Y. The Prohorov distance between any two $l, l' \in \Delta Y$ is equal to the infimum of positive ϵ such that the following inequalities

$$l(\hat{Y}) \leq l'(\hat{Y}^\epsilon) + \epsilon \text{ and } l'(\hat{Y}) \leq l(\hat{Y}^\epsilon) + \epsilon$$

hold for all Borel sets $\hat{Y} \subseteq Y$, where $\hat{Y}^\epsilon = \{y \in Y : \inf_{\hat{y} \in \hat{Y}} d(y, \hat{y}) < \epsilon\}$.

phism $\pi_i : T_i^* \to \Delta(\Theta \times T_{-i}^*)$ associates to each type t_i individual i's belief $\pi_i(t_i)$ over the realized state and other individuals' types. Each type t_i of individual i corresponds in fact to an infinite hierarchy of coherent beliefs, that is, $t_i = (q_i^1(t_i), q_i^2(t_i), \ldots)$, where:

1. Type t_i's *first-order belief* $q_i^1(t_i) \in \Delta\Theta$ is the marginal distribution of $\pi_i(t_i)$ on Θ, describing i's belief regarding the realized state.

2. Type t_i's *second-order belief* $q_i^2(t_i) \in \Delta(\Theta \times (\Delta\Theta)^{I-1})$ describes i's belief regarding the realized state and other individuals' first-order beliefs. It is thus given by:

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

   for all measurable $E \subseteq \Theta \times (\Delta\Theta)^{I-1}$. Notice that $q_i^2(t_i)$ is coherent with $q_i^1(t_i)$ in the sense that the marginal of $q_i^2(t_i)$ on Θ equals $q_i^1(t_i)$.

3. Type t_i's *k^{th}-order belief* $q_i^k(t_i)$ describes i's belief regarding the realized state and up to $(k-1)$ orders of beliefs of other individuals, and is constructed similarly by induction on k.

To simplify notation, $(q_i^1(t_i))_{i \in I}$ will be denoted $q^1(t)$, and $(q_j^1(t_j))_{j \neq i}$ will be denoted $q_{-i}^1(t_{-i})$. A sequence of types $(t_i^n)_{n \geq 1}$ converges to t_i if for each $k \geq 1$, the sequence of k^{th}-order beliefs $(q_i^k(t_i^n))_{n \geq 1}$ converges to $q_i^k(t_i)$ in the weak* topology. Since π_i is a homeomorphism, an equivalent definition is that $(t_i^n)_{n \geq 1}$ converges to t_i if $\pi_i(t_i^n)$ converges to $\pi_i(t_i)$ in the weak* topology.

In applications, one often imposes further restrictions. For instance, depending on circumstances, one may require individuals to be selfish, types to be independent, values to be private, information to be complete, or higher-order beliefs to be derived by Bayes' rule from a common prior defined on states. Each such case can be thought of as restricting attention to a subset $T \subset T^*$ of types, where $T^* = \times_{i \in I} T_i^*$.

Let T_i be the projection of T on T_i^*, that is, the set of types $t_i \in T_i^*$ such that $t \in T$ for some $t_{-i} \in T_{-i}^*$. Clearly, T is a subset of $T_1 \times \cdots \times T_I$. The set T is *belief-closed* if each individual's belief supports only type profiles in T, that

is, $\pi_i(t_i)(\{(\theta, t_{-i}) : (t_i, t_{-i}) \in T\}) = 1$, for all $t_i \in T_i$. This guarantees that the model is common knowledge among individuals. The set T is *regular* if it is belief-closed and the set $Q_i^1(T) = \{q_i^1(t_i) \in \Delta\Theta : t_i \in T_i\}$ of first-order beliefs associated to types in T_i is closed for each individual i. In what follows, it will be useful to distinguish between the product set $\times_{i\in I} Q_i^1(T)$ and the projection of T onto the set of first-order beliefs $Q^1(T) = \{q^1(t) \in (\Delta\Theta)^I : t \in T\}$. Clearly, $Q^1(T) \subseteq \times_{i\in I} Q_i^1(T)$, and the two sets are equal if $Q^1(T)$ has a product structure.

3.3 Social Choice Rules and Simple Direct Mechanisms

The planner's objective is to implement a *social choice function* (SCF) $f : T \to \Delta X$ defined on a regular subset T of T^*, meaning that he wants outcome $f(t)$ to prevail at each $t \in T$.

To achieve this goal, he constructs a *simple direct mechanism* defined on T, which is a measurable function $\mu : M_1 \times \cdots \times M_I \to \Delta X$, where the set M_i of messages is restricted to be $Q_i^1(T)$. Here, 'direct' means that individuals' messages in the mechanism concern only their types, and 'simple' means that the planner bases his decision only on reports about first-order beliefs of individuals.

3.4 Cognitive States

To describe how individuals with bounded depth of reasoning might play a simple direct mechanism μ defined on T, we introduce an individual's cognitive state as in Strzalecki (2010). An individual's cognitive state specifies his depth of reasoning and his belief regarding other individuals' cognitive states. In particular, if an individual's cognitive state is of depth $k \geq 1$, then he believes that every other individual's cognitive state is of at most depth $k-1$. Our only departure from Strzalecki (2010) is the added assumption that individuals of cognitive state of depth 0 play the truth-telling strategy in any simple direct mechanism.

Formally, let $C_i^0 = \{c_i^0\}$ be the singleton set where c_i^0 represents i's cognitive

state of depth 0. Suppose that we have defined the set of cognitive states of depth k', denoted by $C_j^{k'}$, for all individuals $j \in I$ and for all nonnegative integers k' strictly smaller than some $k \geq 1$. Then, individual i's cognitive state of depth k, denoted by c_i^k, is a probability measure over $\cup_{k'=0}^{k-1}(\times_{j\neq i}C_j^{k'})$. Letting $\times_{j\neq i}C_j^{k'} = C_{-i}^{k'}$, we have that $C_i^k = \Delta(\cup_{k'=0}^{k-1}C_{-i}^{k'})$ is the set of individual i's cognitive states of depth k. Note that C_i^1 is a singleton, and C_i^k is compact and metrizable for all $k \geq 0$.

Given the simple direct mechanism μ on T, let $S_i^k(t_i, c_i^k)$ be the set of messages that individual i of type t_i may send when his cognitive state is c_i^k. Formally, these sets are defined by induction on k: $S_i^0(t_i, c_i^0) = \{q_i^1(t_i)\}$, and for each $k > 0$, $m_i \in S_i^k(t_i, c_i^k)$ if

$$m_i \in \arg\max_{m_i' \in M_i} \int_{\Theta \times T_{-i} \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times M_{-i}} U_i(\mu(m_i', m_{-i}), \theta) d\gamma \tag{1}$$

for some conjecture $\gamma \in \Delta(\Theta \times T_{-i} \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times M_{-i})$ such that *(a)* the distribution $\pi_i(t_i)$ coincides with the marginal distribution of γ on $\Theta \times T_{-i}$, *(b)* the distribution c_i^k coincides with the marginal distribution of γ on $\cup_{k'=0}^{k-1} C_{-i}^{k'}$, and *(c)* the marginal distribution of γ on $T_{-i} \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times M_{-i}$ supports a subset of $\cup_{k'=0}^{k-1} \left(\times_{j\neq i} Gr(S_j^{k'})\right)$, where $Gr(S_j^{k'})$ is the graph of $S_j^{k'}$. The conjecture γ represents i's belief regarding the exogenous uncertainty – the state, others' types, and their cognitive states – and endogenous uncertainty – others' messages – he is facing. This belief must be consistent with his belief regarding the state and others' types $\pi_i(t_i)$, his belief regarding others' cognitive states c_i^k, and other players' behavior up to level $k-1$ as captured by conditions *(a)*, *(b)* and *(c)*, respectively. Given his conjecture γ, individual i of type t_i sends a message in order to maximize the expected utility in (1). Let then $\Sigma_i^k(t_i)$ be the set of messages that could be sent by an individual i of type t_i with a depth of reasoning k, that is, $\Sigma_i^k(t_i) = \cup_{c_i^k \in C_i^k} S_i^k(t_i, c_i^k)$. It will be assumed throughout the paper that each individual i's depth of reasoning is bounded by some strictly positive integer K_i. Let then K be the vector $(K_1, \ldots, K_I)$ and $\Sigma_i(t_i) = \cup_{k=0}^{K_i} \Sigma_i^k(t_i)$.

4 Mechanism Design with Bounded Depth of Reasoning

The simple direct mechanism μ on T *implements* the SCF $f : T \to \Delta X$ *when individuals' depth of reasoning is bounded by* K if the following two conditions are satisfied:

1. $S_i^k(t_i, c_i^k) \neq \emptyset$ for all $t_i \in T_i$, $c_i^k \in C_i^k$, $0 \leq k \leq K_i$, and $i \in I$.
2. For each $t \in T$, if $m_i \in \Sigma_i(t_i)$ for each i, then $\mu(m_1, \ldots, m_I) = f(t)$.

The SCF f is then said to be *implementable when individuals' depth of reasoning is bounded by* K.

Being unsure about the individuals' cognitive states, we thus require in each information state that (1) any cognitive state admits at least one message that is consistent with it, and (2) the mechanism delivers the desired outcome for all message profiles that are consistent with at least one combination of cognitive states. Implementation in this sense is quite flexible, as cognitive states accomodate a variety of reasonings (and thus behaviors), including for instance the "cognitive hierarchy" model of Stahl (1993) (see also Stahl and Wilson (1995) and Camerer *et al.* (2004)), or the "level k" model used by Costa-Gomes and Crawford (2006) and others. While related to rationalizable full implementation, also with an iterative construction, our definition is less demanding, as individuals' depth of reasoning is bounded and all cognitive states start with truth-telling. As we now show, implementation with bounded depth of reasoning is closely related to (interim Bayesian) incentive compatibility.

The simple direct mechanism μ defined on T is *incentive compatible* if

$$\int_{\Theta \times T_{-i}} U_i(\mu(q^1(t)), \theta) d\pi_i(t_i) \geq \int_{\Theta \times T_{-i}} U_i(\mu(m_i, q_{-i}^1(t_{-i})), \theta) d\pi_i(t_i),$$

for all $m_i \in M_i$, $t_i \in T_i$ and $i \in I$ (recall that $M_i = Q_i^1(T)$, and so the above inequality means that each type of each player wants to report his true first-order belief when everyone else reports their first-order beliefs truthfully). It is *strictly incentive compatible* if these inequalities are strict for all $m_i \neq q_i^1(t_i)$.

The mechanism μ *achieves an SCF* $f : T \to \Delta X$ if it generates f when individuals are truth-telling: $\mu(q^1(t)) = f(t)$, for all $t \in T$.[6]

Theorem 1. *(a) If an SCF is implementable when individuals' depth of reasoning is bounded by K, then it can be achieved by a simple direct mechanism that is incentive compatible.*

(b) If an SCF can be achieved by a simple direct mechanism that is strictly incentive compatible, then it is implementable when individuals' depth of reasoning is bounded by K.

Proof. (a) Let μ be a simple direct mechanism that implements the SCF f when individuals' depth of reasoning is bounded. The fact that μ achieves f follows from the second condition in the definition of implementability given that $q_i^1(t_i) \in \Sigma_i^0(t_i)$ for each t_i and i. The mechanism μ must also be incentive compatible. Otherwise, there exists i, t_i, m_i such that

$$\int_{\Theta \times T_{-i}} U_i(\mu(q^1(t)), \theta) d\pi_i(t_i) < \int_{\Theta \times T_{-i}} U_i(\mu(m_i, q_{-i}^1(t_{-i})), \theta) d\pi_i(t_i). \quad (2)$$

By the first condition of implementability, let $m_i^* \in \Sigma_i^1(t_i)$. By the second condition of implementability, $f(t_i, t_{-i}) = \mu(m_i^*, q_{-i}^1(t_{-i}))$, for all t_{-i} such that $(t_i, t_{-i}) \in T$. Since T is belief-closed, we have

$$\int_{\Theta \times T_{-i}} U_i(f(t), \theta) d\pi_i(t_i) = \int_{\Theta \times T_{-i}} U_i(\mu(m_i^*, q_{-i}^1(t_{-i})), \theta) d\pi_i(t_i). \quad (3)$$

We have thus reached a contradiction: the left-hand sides of (2) and (3) coincide, since μ achieves f, and the right-hand-side of (3) is at least as large as that of (2), since m_i^* is a best-response to truth-telling.

(b) Suppose that $f : T \to \Delta X$ is achieved by a simple direct mechanism μ that is strictly incentive compatible. It is then easy to check by induction on k that $S_i^k(t_i, c_i^k) = \{q_i^1(t_i)\}$ for all t_i, c_i^k, k, and i, as the *unique* best response

[6] Hence, SCFs that can be achieved through simple direct mechanisms are invariant to second- or higher-order beliefs.

to truth-telling is telling the truth. Hence μ implements f when individuals' depth of reasoning is bounded. $\square$

On the one hand, although strict incentive compatibility is sufficient, it is not necessary for implementation when individuals' depth of reasoning is bounded. This is easily illustrated in the following example:

Example 1. Let $I = \{1, 2\}$, $\Theta = \{\theta, \theta'\}$, and $X = \{a, b, c\}$. The Bernoulli utility functions of the agents are such that a is the worst alternative for agent 1 in state θ – i.e., $u_1(a, \theta) < \min\{u_1(b, \theta), u_1(c, \theta)\}$ – whereas a is the best alternative for agent 2 in state θ – i.e., $u_2(a, \theta) > \max\{u_2(b, \theta), u_2(c, \theta)\}$. Lastly, suppose b is the best alternative for both agents in state θ'.

Consider the complete information model in which the state is common knowledge. Let t_i^θ and $t_i^{\theta'}$ denote the complete information types of player i associated with states θ and θ', respectively. Let $T = \{(t_1^\theta, t_2^\theta), (t_1^{\theta'}, t_2^{\theta'})\}$, and suppose the SCF $f : T \to \Delta X$ is such that $f(t_1^\theta, t_2^\theta) = a$ and $f(t_1^{\theta'}, t_2^{\theta'}) = b$. It is straightforward to argue that the following simple direct mechanism implements f when individuals' depth of reasoning is bounded:

	$q_2^1(t_2^\theta)$	$q_2^1(t_2^{\theta'})$
$q_1^1(t_1^\theta)$	a	c
$q_1^1(t_1^{\theta'})$	a	b

Notice that the above mechanism is not strictly incentive compatible. In fact, since a is the worst alternative for type t_1^θ, there does not exist any strictly incentive compatible simple direct mechanism that implements f when individuals' depth of reasoning is bounded. $\diamond$

On the other hand, although incentive compatibility is necessary, it is never sufficient just by itself for implementation when individuals' depth of reasoning is bounded. To be precise, if an incentive compatible simple direct mechanism μ implements f when individuals' depth of reasoning is bounded, then μ must either be strictly incentive compatible or satisfy the following condition: For all $t \in T$ and $t' \neq t$ such that for each player i, type t_i is indifferent between $q_i^1(t_i)$ and $q_i^1(t_i')$ when others play their truth-telling strategies in μ, we must

have $\mu(q^1(t')) = \mu(q^1(t)) = f(t)$ because $q_i^1(t_i') \in \Sigma_i^1(t_i), \forall t_i$.[7] Indeed, when individuals' depth of reasoning may be greater than 1, then best responses to beliefs that support message profiles that are in turn best responses to truth-telling must also achieve f, and so on. However, we do not present these additional necessary requirements as they are not relevant for what follows.

5 Robustness to Small Modeling Mistakes

As pointed out by Börgers and Oh (2012), applied game theory often focuses on naive type spaces where two different types of an individual corresponds to two different preference orderings. Mechanism design is no exception. Most often, additional restrictions are imposed, including for instance that individuals are selfish, and/or that individuals' payoff irrelevant beliefs do not vary with their types, and/or that individuals' beliefs are independent, etc.

It is natural then to ask whether results that hold for small type spaces are robust against possible modeling misspecifications. Even if one is overall confident that information and beliefs are aptly described by the subset of types $\hat{T} \subseteq T^*$, it would be preferable that we have a mechanism that does not implement dramatically different outcomes when considering any nearby type profile in T^*. We assume throughout the section that $\hat{T}$ is regular.

Consider an SCF $f : \hat{T} \to \Delta X$. A simple direct mechanism μ defined on T^* *continuously implements* f *when individuals' depth of reasoning is bounded by* K if the following two conditions are satisfied:

1. $S_i^k(t_i, c_i^k) \neq \emptyset$ for all $t_i \in T_i^*$, $c_i^k \in C_i^k$, $0 \leq k \leq K_i$, and $i \in I$.
2. For each sequence $(t^n)_{n\geq 1}$ in T^* that converges to some $\hat{t} \in \hat{T}$, if $m_i^n \in \Sigma_i(t_i^n)$ for each i and each n, then $(\mu(m_1^n, \ldots, m_I^n))_{n\geq 1}$ converges to $f(\hat{t})$.

The SCF f is then said to be *continuously implementable when individuals' depth of reasoning is bounded by* K.

Theorem 2. *(a) Suppose that the SCF $f : \hat{T} \to \Delta X$ is continuously implementable when individuals' depth of reasoning is bounded by K, then it can*

[7] Note that if $t' \in T$, then this further implies that $f(t') = f(t)$.

be achieved by a simple direct mechanism defined on $\hat{T}$ that is incentive compatible and continuous at all first-order belief profiles in $Q^1(\hat{T})$.

(b) Suppose $f : \hat{T} \to \Delta X$ is achievable through a simple direct mechanism defined on $\hat{T}$ that is both strictly incentive compatible and continuous. Then f is continuously implementable when individuals' depth of reasoning is bounded by K.

The necessary condition for continuous implementability says that the simple direct mechanism defined on $\hat{T}$ that achieves f must be continuous at all points in $Q^1(\hat{T})$, which is the projection of $\hat{T}$ onto the set of first-order beliefs. In contrast, the sufficient condition requires the simple direct mechanism defined on $\hat{T}$ to be continuous at all points in its domain $\times_{i\in I} Q_i^1(\hat{T})$. The stronger continuity requirement in the sufficient condition is trivially satisfied when $Q_i^1(\hat{T})$ is finite for all i, which will be the case whenever $\hat{T}$ is finite (as in the motivating example in Section 2). There is again no gap between the necessary and sufficient continuity requirements when $Q^1(\hat{T})$ is itself a product set, which will be the case whenever $\hat{T}$ has a product structure. Although several interesting applications have $Q^1(\hat{T})$ as a product set (e.g., the bilateral trading example in the next section), other applications do not (e.g., complete information type spaces).

Proof. (a) Suppose the simple direct mechanism μ on T^* continuously implements $f : \hat{T} \to \Delta X$ when individuals' depth of reasoning is bounded by K. The domain of μ equals $(\Delta\Theta)^I$. Define $\hat{\mu}$ as the restriction of μ to $\times_{i\in I} Q_i^1(\hat{T})$, that is, $\hat{\mu} : \times_{i\in I} Q_i^1(\hat{T}) \to \Delta X$ such that $\hat{\mu}(m_1, \ldots, m_I) = \mu(m_1, \ldots, m_I), \forall (m_1, \ldots, m_I) \in \times_{i\in I} Q_i^1(\hat{T})$. Thus defined, $\hat{\mu}$ is a simple direct mechanism on $\hat{T}$.

Pick any $\hat{t} \in \hat{T}$. In μ, we have $q_i^1(\hat{t}_i) \in \Sigma_i^0(\hat{t}_i), \forall i$. Then $\mu(q^1(\hat{t})) = f(\hat{t})$ by the second condition of continuous implementability (use the constant sequence of types fixed at $\hat{t}$). Since $\hat{\mu}(q^1(\hat{t})) = \mu(q^1(\hat{t}))$, the mechanism $\hat{\mu}$ achieves f.

The mechanism $\hat{\mu}$ must also be incentive compatible. Otherwise, there

exists $i, \hat{t}_i \in \hat{T}_i$, and $m_i \in Q_i^1(\hat{T})$ such that

$$\int_{\Theta \times \hat{T}_{-i}} U_i(\hat{\mu}(q_i^1(\hat{t}_i), q_{-i}^1(t_{-i})), \theta) d\pi_i(\hat{t}_i) < \int_{\Theta \times \hat{T}_{-i}} U_i(\hat{\mu}(m_i, q_{-i}^1(t_{-i})), \theta) d\pi_i(\hat{t}_i). \tag{4}$$

Since $\hat{T}$ is belief-closed and $\hat{\mu}$ is a restriction of μ to $\times_{i \in I} Q_i^1(\hat{T})$, (4) is equivalent to

$$\int_{\Theta \times T_{-i}^*} U_i(\mu(q_i^1(\hat{t}_i), q_{-i}^1(t_{-i})), \theta) d\pi_i(\hat{t}_i) < \int_{\Theta \times T_{-i}^*} U_i(\mu(m_i, q_{-i}^1(t_{-i})), \theta) d\pi_i(\hat{t}_i). \tag{5}$$

By the first condition of continuous implementability, let $m_i^* \in \Sigma_i^1(\hat{t}_i)$ in the mechanism μ. By the second condition of continuous implementability, $\mu(m_i^*, q_{-i}^1(t_{-i})) = f(\hat{t}_i, t_{-i}) = \mu(q_i^1(\hat{t}_i), q_{-i}^1(t_{-i}))$, for all t_{-i} such that $(\hat{t}_i, t_{-i}) \in \hat{T}$. Since $\hat{T}$ is belief-closed, we have

$$\int_{\Theta \times T_{-i}^*} U_i(\mu(q_i^1(\hat{t}_i), q_{-i}^1(t_{-i})), \theta) d\pi_i(\hat{t}_i) = \int_{\Theta \times T_{-i}^*} U_i(\mu(m_i^*, q_{-i}^1(t_{-i})), \theta) d\pi_i(\hat{t}_i). \tag{6}$$

We have thus reached a contradiction: the right-hand side of (6) is at least as large as that of (5), since m_i^* is a best-response to truth-telling.

Finally, pick any $\hat{t} \in \hat{T}$ and $q^1(\hat{t})$. Consider any sequence $(m^n)_{n \geq 1}$ of first-order beliefs in $\times_{i \in I} Q_i^1(\hat{T})$ that converges to $q^1(\hat{t})$. Let $(t^n)_{n \geq 1}$ be any sequence of types in T^* that converges to $\hat{t}$ such that $q^1(t^n) = m^n, \forall n$. In mechanism μ, we have $q_i^1(t_i^n) \in \Sigma_i^0(t_i^n)$ for all i and n. Hence, by the second condition of continuous implementability, $\mu(q^1(t^n)) = \mu(m^n) = \hat{\mu}(m^n)$ converges to $f(\hat{t}) = \mu(q^1(\hat{t})) = \hat{\mu}(q^1(\hat{t}))$.

(b) Let $\hat{\mu} : \times_{i \in I} Q_i^1(\hat{T}) \to \Delta X$ be the simple direct mechanism that achieves f in the statement. To prove that f is continuously implementable, we must propose a mechanism defined for unrestricted message profiles, that is, whose domain is $(\Delta\Theta)^I$ instead of $\times_{i \in I} Q_i^1(\hat{T})$. The strategy of proof is to apply $\hat{\mu}$ after 'translating' messages in $\Delta\Theta \setminus Q_i^1(\hat{T})$ into messages in $Q_i^1(\hat{T})$, keeping messages in $Q_i^1(\hat{T})$ unchanged.[8] The following lemma is a variant of Dugundji

[8] An obvious choice would be to use a single-valued selection of the projection operator

(1951).

Lemma 1. *For each $i \in I$, there exists a correspondence $\omega_i : \Delta\Theta \to Q_i^1(\hat{T})$ with nonempty finite values and for each message $m_i \in \Delta\Theta$, there exists a probability distribution ξ_{m_i} with full support on $\omega_i(m_i)$ such that $\mu : (\Delta\Theta)^I \to \Delta X$ extends $\hat{\mu}$ continuously, where μ is the mechanism that associates to any message profile $m \in (\Delta\Theta)^I$ the lottery that selects $\hat{\mu}(q^1)$ with probability $\times_{i\in I}\xi_{m_i}(q_i^1)$, for all $q^1 \in \times_{i\in I}\omega_i(m_i)$.*

The mechanism μ thus 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 $\xi_{m_i}(q_i^1)$. The detailed construction of the translations can be found in the Appendix.[9]

The Appendix also contains the proofs of the following two lemmas. First, if i's type $\hat{t}_i$ belongs to the restricted domain $\hat{T}_i$, then for any bounded depth of reasoning his report m_i under μ will be such that its translation is $q_i^1(\hat{t}_i)$ with probability 1, that is, $\omega_i(m_i) = \{q_i^1(\hat{t}_i)\}$. Second, given any bounded depth of reasoning, an individual's set of strategies compatible with such depth is upper hemicontinuous in his type.

Lemma 2. *For all $i \in I$ and $k \geq 0$, the correspondence Σ_i^k in μ is such that $\omega_i(m_i) = \{q_i^1(\hat{t}_i)\}$, for each $m_i \in \Sigma_i^k(\hat{t}_i)$ and $\hat{t}_i \in \hat{T}_i$.*

Lemma 3. *For all $i \in I$ and $k \geq 0$, the correspondence Σ_i^k in μ is upper hemicontinuous.*

We are now ready to prove that μ continuously implements f.

---

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 a more elaborate construction based on 'probabilistic translations,' as we do.

[9]For such a step, one could take a host of alternative approaches. For example, one could apply Dugundji's result as is if we overlook the product structure, or one could apply his result component by component. We find it more convenient to provide a new construction. 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 the input m_i. This kind of product/separability property is needed in Lemma 2 of the proof.

To begin with, for each i and $k \geq 0$, the correspondence S_i^k has nonempty values. This follows by assumption when $k = 0$ whereas when $k \geq 1$, then for each $t_i \in T_i^*$ and $c_i^k \in C_i^k$, best responses to any consistent conjecture γ must be nonempty since the objective function is continuous – as both U_i and μ are continuous – and the set of messages $\Delta\Theta$ is compact (see the proof of Lemma 3 for a detailed argument that establishes that the objective function is continuous).

To finish the proof, let $(t^n)_{n\geq 1}$ be a sequence of type profiles converging to some $\hat{t} \in \hat{T}$. For each i and n, pick any $m_i^n \in \Sigma_i(t_i^n)$. We show that $\mu((m_i^n)_{i\in I})$ converges to $f(\hat{t})$ – we want to show this even though $(m_i^n)_{i\in I}$ may not be convergent, which makes the following argument slightly longer than one would have expected. Compactness of $\Delta\Theta$ implies that every *subsequence* of $(m_i^n)_{i\in I}$ has a subsequence $(m_i^{n_l})_{i\in I}$ that converges to some message profile m. By Lemma 3, the correspondence $\Sigma_i = \cup_{k=0}^{K_i} \Sigma_i^k$ is upper hemicontinuous, and hence $m_i \in \Sigma_i(\hat{t}_i)$. So $\omega_i(m_i) = \{q_i^1(\hat{t}_i)\}$, by Lemma 2. Since μ is continuous, $\mu((m_i^{n_l})_{i\in I})$ must converge to $\mu(m) = \hat{\mu}(q^1(\hat{t})) = f(\hat{t})$. This argument implies that every *subsequence* of $\mu((m_i^n)_{i\in I})$ has a subsequence that converges to $f(\hat{t})$, which is sufficient to conclude that the sequence $\mu((m_i^n)_{i\in I})$ itself converges to $f(\hat{t})$. □

6 Examples

In order to understand (weak) Bayesian implementability, much effort has been devoted over the years to identify mechanisms that are incentive compatible. Fortunately, we can build on this work to understand implementability under bounded depth of reasoning, as it is guaranteed under a similar condition (see Section 4). Though similar, it is nevertheless a bit stronger, as incentive constraints must be satisfied strictly in our sufficient condition. Perhaps what is even more surprising, in view of the difficulty of achieving continuous implementation in Bayesian Nash equilibrium, is that when individuals' depth of reasoning is bounded, continuous implementation obtains as soon as the mechanism implementing the SCF is also continuous (see Section 5).

Continuity of the mechanism is automatically satisfied when the initial set of types $\hat{T}$ or, more generally, each player's first-order belief in $\hat{T}$ is finite. In these cases, we only need to check strict incentive compatibility of the simple direct mechanism on $\hat{T}$ to guarantee continuous implementation under bounded depth of reasoning. For instance, the SCF in our motivating example is continuously implementable under bounded depth of reasoning. This section illustrates with a few classic applications how requiring continuity and strict incentive compatibility is not much more demanding than imposing standard incentive constraints. Investigating the properties of continuity and strict incentive compatibility more systematically is an interesting research agenda for the future.

The first example shows that the classic expected externality mechanism (see d'Aspremont and Gerard-Varet (1979)) does guarantee continuity and strict incentive compatibility in a large class of public good problems.

Example 2 (Public Good Decision)**.** Consider a public good problem with quasilinear utilities. The public decision to be implemented belongs to a compact convex metric space A, individual i's payoff type θ_i belongs to a compact metric space Θ_i, and utility functions for the public decision are given by

$$u_i(a, \theta) = v_i(a, \theta_i) + w_i(a, \theta_{-i}) + y(a, \theta),$$

for each $a \in A$ and each state $\theta \in \Theta = \times_{i \in I} \Theta_i$. In addition to the public decision, the mechanism may impose a monetary transfer $z_i \in [-z^*, z^*]$ on individual i. The total utility for individual i when a is implemented while receiving a net transfer z_i is equal to $u_i(a, \theta) + z_i$, for all states θ. This general description contains the classic case of private values (with $w_i = y = 0$). More generally, we also allow for other players' payoff types to impact player i's utility, either in a way that is additively separable and/or through a general common interest term y.

The planner is interested in a regular subset of types $\hat{T}$ in which it is common knowledge that each individual knows his own payoff type, and for each $i \in I$ and each $\theta_i \in \Theta_i$, there exists $\hat{t}_i \in \hat{T}_i$ such that type $\hat{t}_i$'s payoff

type $\theta_i(\hat{t}_i) = \theta_i$. Thus the first-order belief of any type $\hat{t}_i \in \hat{T}_i$ specifies that his payoff type equals $\theta_i(\hat{t}_i)$ and his belief regarding other individuals' payoff types.

Consider now the following decision rule

$$a(\theta) = \arg\max_{a \in A} [y(a, \theta) + \sum_{i \in I} v_i(a, \theta_i)], \ \forall \theta,$$

and the following transfers (assuming z^* is sufficiently large to allow them)

$$z_i(\theta) = -w_i(f(\theta), \theta_{-i}) + \sum_{j \in I \setminus \{i\}} v_j(f(\theta), \theta_j), \ \forall \theta.$$

When values are private ($w_i = y = 0$ for all i), $a(\cdot)$ picks decisions that are ex-post efficient (maximizing the utilitarian objective), and our definition boils down to d'Aspremont and Gerard-Varet's expected externality mechanism.

We now show that continuity and strict incentive compatibility obtain in a large class of problems of this type, namely whenever (a) $y(a, \theta) + \sum_{i \in I} v_i(a, \theta_i)$ is strictly concave in a and (b) all the v_i's, w_i's, and y are continuous in both arguments. Indeed, continuity of the mechanism then follows from Berge's Maximum Theorem. As for strict incentive compatibility, observe that individual i of type $\hat{t}_i$ chooses his report θ_i' to maximize the following expression:

$$\int_{\Theta \times \hat{T}_{-i}} [u_i\big(a(\theta_i', \theta_{-i}(t_{-i})), (\theta_i(\hat{t}_i), \theta_{-i}(t_{-i}))\big) + z_i(\theta_i', \theta_{-i}(t_{-i}))] d\pi_i(\hat{t}_i),$$

which amounts to

$$\int_{\Theta \times \hat{T}_{-i}} [y\big(a(\theta_i', \theta_{-i}(t_{-i})), (\theta_i(\hat{t}_i), \theta_{-i}(t_{-i}))\big) + \sum_{j \in I} v_j\big(f(\theta_i', \theta_{-i}(t_{-i})), \theta_j(t_j)\big)] d\pi_i(\hat{t}_i).$$

The mechanism $(a(\cdot), (z_i(\cdot))_{i \in I})$ is thus strictly incentive compatible.[10] $\diamond$

[10]The mechanism $(a(\cdot), (z_i(\cdot))_{i \in I})$ is such that each individual reports only his payoff type. In contrast, a simple direct mechanism defined on $\hat{T}$ is such that each individual reports his payoff type *and* his belief regarding other individuals' payoff types. Thus $(a(\cdot), (z_i(\cdot))_{i \in I})$ is equivalent to a simple direct mechanism $\hat{\mu}$ defined on $\hat{T}$ that is unresponsive to individuals' reports about their beliefs regarding other individuals' payoff types, i.e., $\hat{\mu}(q^1(t)) = (a(\theta(t)), (z_i(\theta(t)))_{i \in I}), \forall q^1(t) \in \times_{i \in I} Q_i^1(\hat{T})$. Continuity of $(a(\cdot), (z_i(\cdot))_{i \in I})$ implies continuity of $\hat{\mu}$. However, strict incentive compatibility of $(a(\cdot), (z_i(\cdot))_{i \in I})$ does not

The next example investigates a large class of bilateral trade problems with independent private values, as in Myerson and Satterthwaite (1983). While the second-best mechanism they identify is discontinuous and only weakly incentive compatible, we show that, if the inverse hazard rates associated with the distributions of buyer's value and seller's cost are increasing, then we can continuously implement an approximately optimal SCF for any bounded depth of reasoning.

Example 3 (Bilateral Trading)**.** There are two traders of an indivisible object, buyer b and seller s. A state is a pair (v, c), where $v \in V = [0, 1]$ is the buyer's value and $c \in C = [0, 1]$ is the seller's cost. The planner is interested in the set $\hat{T}$ of types of the traders such that it is common knowledge that each trader knows his value/cost and that trader i's value/cost is distributed on $[0, 1]$ according to G_i with continuous and positive density g_i. Any two types of the buyer (seller) differ only in their values (costs) since other beliefs in the infinite hierarchy of beliefs are pinned down by the common knowledge assumption. Hence, instead of explicitly describing each type as an infinite hierarchy of beliefs, we use the equivalent implicit formulation with $\hat{T} = \hat{T}_b \times \hat{T}_s$, where for each trader i, the set of his types $\hat{T}_i = [0, 1]$, and his belief $\pi_i : \hat{T}_i \to \Delta(V \times C \times \hat{T}_j)$ is given as follows: The buyer (*resp.* seller) of type t_b (*resp.* t_s) knows that his value (*resp.* cost) equals t_b (*resp.* t_s) and believes that the

imply incentive compatibility of $\hat{\mu}$. Indeed, since $\hat{\mu}$ is unresponsive to reports about beliefs regarding other individuals' payoff types, individuals' incentives are unaffected by such reports. Nevertheless, since $(a(\cdot), (z_i(\cdot))_{i \in I})$ is strictly incentive compatible, in mechanism $\hat{\mu}$, each individual has the strict incentive to report his payoff type truthfully when others report their payoff types truthfully irrespective of their reports about their beliefs regarding other individuals' payoff types. This property implies that when we continuously extend $\hat{\mu}$ to $\Delta\Theta$ as in the proof of Theorem 2b, for all i and k, the correspondence Σ_i^k in the extended mechanism μ will be such that for each $\hat{t}_i \in \hat{T}_i$ and $m_i \in \Sigma_i^k(\hat{t}_i)$, the translation $\omega_i(m_i)$ will equal the set of all those first-order beliefs in $Q_i^1(\hat{T})$ under which individual i's payoff type equals $\theta_i(\hat{t}_i)$. With this change in the statement of Lemma 2 – which is the only place where we used the strict incentive compatibility of the simple direct mechanism on $\hat{T}$ in the original proof –, the rest of the argument of Theorem 2b can be replicated to show that μ will continuously implement the outcome $(a(\cdot), (z_i(\cdot))_{i \in I})$. More generally, when $\hat{T}$ is "known own payoff" type space, then any decision rule defined on the set of payoff types Θ can be continuously implemented under bounded depth of reasoning as long as it is strictly incentive compatible and continuous over Θ.

seller's cost (*resp.* buyer's value) equals his type which is distributed according to G_s (*resp.* G_b).

A (nonrandom) alternative specifies $p \in \{0,1\}$, where $p = 1$ means that the object 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.

Consider any simple direct mechanism $\hat{\mu} : Q_b^1(\hat{T}) \times Q_s^1(\hat{T}) \to \Delta X$. The first-order belief of type t_i of trader i is that his value/cost equals t_i and that the opponent j's value/cost is distributed according to G_j. Since the belief regarding the opponent's value/cost is independent of the trader's value/cost, the simple direct mechanism $\hat{\mu}$ is equivalent to a direct mechanism $\nu : \hat{T}_b \times \hat{T}_s \to \Delta X$ such that

$$\nu(t_b, t_s) = \hat{\mu}(q_b^1(t_b), q_s^1(t_s)), \forall (t_b, t_s) \in \hat{T}_b \times \hat{T}_s.$$

We therefore now work with direct mechanisms.

Given a direct mechanism ν, it is straightforward to define 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 $(t_b, t_s) \in \hat{T}_b \times \hat{T}_s$. Then, let $p_i(t_i) = \int_0^1 p(t_i, t_j) g_j(t_j) dt_j$ be the probability that trader i of type t_i trades and $z_i(t_i) = \int_0^1 z(t_i, t_j) g_j(t_j) dt_j$ be his expected transfer. If the buyer of type t_b reports t_b' in the direct mechanism ν, then his expected payoff is $t_b p_b(t_b') - z_b(t_b')$. If the seller of type t_s reports t_s' in the direct mechanism ν, then his expected payoff is $z_s(t_s') - t_s p_s(t_s')$.

The following result follows from Theorem 2 in Myerson and Satterthwaite (1983): If $\frac{G_b(t_s)-1}{g_b(t_b)}$ and $\frac{G_s(t_s)}{g_s(t_s)}$ are strictly increasing functions on $[0,1]$, then there exists an incentive compatible and interim individually rational direct mechanism ν^* that maximizes the ex-ante gains from trade. Furthermore, the probability of trade function corresponding to ν^* is such that there exists an $\alpha \in (0,1)$ such that

$$p^*(t_b, t_s) = \begin{cases} 1, & \text{if } t_b - t_s \geq \alpha \left(\frac{1-G_b(t_b)}{g_b(t_b)} + \frac{G_s(t_s)}{g_s(t_s)} \right) \\ 0, & \text{if } t_b - t_s < \alpha \left(\frac{1-G_b(t_b)}{g_b(t_b)} + \frac{G_s(t_s)}{g_s(t_s)} \right). \end{cases}$$

This optimal ν^* is not continuous because the corresponding probability

of trade function $p^*(.,.)$ is not continuous. However, we now approximate ν^* by a continuous direct mechanism that is strictly incentive compatible and interim individually rational. To do that, pick the associated α, and then for all $\beta \in [\alpha, 1]$ and $l \in [1, \infty) \cup \{\infty\}$, define

$$p^{\beta,l}(t_b, t_s) = \begin{cases} 1, & \text{if } t_b - t_s \geq \beta\left(\frac{1-G_b(t_b)}{g_b(t_b)} + \frac{G_s(t_s)}{g_s(t_s)}\right) \\ \left(\frac{t_b - t_s}{\beta\left(\frac{1-G_b(t_b)}{g_b(t_b)} + \frac{G_s(t_s)}{g_s(t_s)}\right)}\right)^l, & \text{if } 0 < t_b - t_s < \beta\left(\frac{1-G_b(t_b)}{g_b(t_b)} + \frac{G_s(t_s)}{g_s(t_s)}\right) \\ 0, & \text{if } t_b - t_s \leq 0, \end{cases}$$

where we let $\left(\frac{t_b - t_s}{\beta\left(\frac{1-G_b(t_b)}{g_b(t_b)} + \frac{G_s(t_s)}{g_s(t_s)}\right)}\right)^\infty = 0$. Thus, $p^{\alpha,\infty}(.,.) = p^*(.,.)$.

Also, define

$$\lambda(\beta, l) = \int_0^1 \int_0^1 \left(t_b + \frac{G_b(t_b) - 1}{g_b(t_b)} - t_s - \frac{G_s(t_s)}{g_s(t_s)}\right) p^{\beta,l}(t_b, t_s) g_b(t_b) g_s(t_s) dt_s dt_b.$$

Since $\frac{G_b(t_s)-1}{g_b(t_b)}$ and $\frac{G_s(t_s)}{g_s(t_s)}$ are strictly increasing and continuous, it is easy to see that for all $(\beta, l) \in [\alpha, 1] \times [1, \infty)$, $p^{\beta,l}(.,.)$ is continuous, $p_b^{\beta,l}(t_b)$ is strictly increasing, and $p_s^{\beta,l}(t_s)$ is strictly decreasing. We can also show that $\lambda(\beta, l)$ is strictly increasing in β and l. Moreover, $\lim_{l\to\infty} \lambda(\beta, l) = \lambda(\beta, \infty), \forall \beta$.

We know from Theorems 1 and 2 in Myerson and Satterthwaite (1983) that $\lambda(\alpha, \infty) = 0$. Since λ is strictly increasing in β, we have $\lambda(\beta, \infty) > 0$ for all $\beta > \alpha$. As a result, for all $\beta > \alpha$, there exists $l(\beta) < \infty$ such that $\lambda(\beta, l) \geq 0$ for all $l \geq l(\beta)$. Then using the construction in Theorem 1 in Myerson and Satterthwaite (1983) for all β and $l \geq l(\beta)$, we can find a continuous expected payment function $z^{\beta,l}(t_b, t_s)$, and hence a continuous direct mechanism $\nu^{\beta,l}$ that is incentive compatible and interim individually rational. In fact, since $p_b^{\beta,l}(t_b)$ is strictly increasing and $p_s^{\beta,l}(t_s)$ is strictly decreasing, $\nu^{\beta,l}$ is strictly incentive compatible. Theorem 2b thus implies that the SCF $\nu^{\beta,l}$ is continuously implementable on $\hat{T}_b \times \hat{T}_s$ when individuals' depth of reasoning is bounded by K.

By taking β close enough to α and l large enough, we can approximate the optimal ν^* by $\nu^{\beta,l}$. Thus, there exist approximately optimal SCFs on $\hat{T}$

that are continuously implementable when individuals' depth of reasoning is bounded. ◇

To conclude this section, we provide an example with multidimensional types as in Jehiel *et al.* (2012). Utility functions are picked so as to satisfy their generic condition where locally robust implementation in their sense is impossible. By contrast, continuous implementation in our sense is feasible.

Example 4. There are two individuals, 1 and 2. A state is a pair (θ_1, θ_2), where $\theta_i = (\theta_{i1}, \theta_{i2})$ is individual i's payoff type drawn from $\Theta_i = [0,1]^2$. The planner is interested in the set $\hat{T}$ of types of the individuals such that it is common knowledge that each individual i knows his payoff type θ_i and that his payoff type is distributed independently and uniformly on Θ_i.

As in the bilateral trading example, we use the implicit formulation of the type space with $\hat{T} = \hat{T}_1 \times \hat{T}_2$, where for each individual i, the set of his types $\hat{T}_i = \Theta_i$, and his belief $\pi_i : \hat{T}_i \to \Delta(\Theta_1 \times \Theta_2 \times \hat{T}_j)$ is given as follows: The individual of type t_i knows that his payoff type equals t_i and believes that individual j's payoff type equals his type which is distributed uniformly on $[0,1]^2$.

There are two possible social decisions, $x \in \{0,1\}$. The planner can impose any monetary transfer $z_i \in [-z^*, z^*]$ on player i, where z^* is sufficiently large. Player i's Bernoulli utility function is $u_i((x,z_i),(\theta_i,\theta_{-i})) = xv_i(\theta_i,\theta_{-i}) - z_i$.

Again, as in the bilateral trading example, instead of simple direct mechanisms, we can work with an allocation rule $p : \Theta_1 \times \Theta_2 \in [0,1]$, where $p(\theta_1,\theta_2)$ is the probability of implementing decision 1 when the individuals' types are (θ_1,θ_2), and a transfer rule $z : \Theta_1 \times \Theta_2 \to [-z^*, z^*]^2$ with $z_i(\theta_1,\theta_2)$ being the monetary transfer imposed on player i when types are (θ_1,θ_2).

Fix (v_1, v_2) to be a pair of generic bilinear value functions as defined by Jehiel *et al.* (2012). For instance,

$$v_1(\theta_1,\theta_2) = (2 + 8\theta_{21} + 9\theta_{22})\theta_{11} + (1 + 4\theta_{21} + 6\theta_{22})\theta_{12} + 3\theta_{21} + 5\theta_{22}$$
$$v_2(\theta_1,\theta_2) = (40 + 16\theta_{11} + 9\theta_{12})\theta_{21} + (14 + 12\theta_{11} + 14\theta_{12})\theta_{22} + \theta_{11} + 2\theta_{12}.$$

Now, consider the following allocation and transfer rules:

$$
\begin{aligned}
p(\theta_1,\theta_2) =& \frac{1}{3}(\theta_{11}+\theta_{21}+\theta_{12}\theta_{22}) \\
z_1(\theta_1,\theta_2) =& v_1(\theta_1,\theta_2)p(\theta_1,\theta_2) - \int_0^{\theta_{11}} p((\tilde{\theta}_{11},\theta_{12}),\theta_2)\frac{\partial v_1((\tilde{\theta}_{11},\theta_{12}),\theta_2)}{\partial\theta_{11}}d\tilde{\theta}_{11} \\
& - \int_0^{\theta_{12}} p((\theta_{11},\tilde{\theta}_{12}),\theta_2)\frac{\partial v_1((\theta_{11},\tilde{\theta}_{12}),\theta_2)}{\partial\theta_{12}}d\tilde{\theta}_{12} + 2\theta_{11}\theta_{12} \\
z_2(\theta_1,\theta_2) =& v_2(\theta_1,\theta_2)p(\theta_1,\theta_2) - \int_0^{\theta_{21}} p(\theta_1,(\tilde{\theta}_{21},\theta_{22}))\frac{\partial v_2(\theta_1,(\tilde{\theta}_{21},\theta_{22}))}{\partial\theta_{21}}d\tilde{\theta}_{21} \\
& - \int_0^{\theta_{22}} p(\theta_1,(\theta_{21},\tilde{\theta}_{22}))\frac{\partial v_2(\theta_1,(\theta_{21},\tilde{\theta}_{22}))}{\partial\theta_{22}}d\tilde{\theta}_{22} + 9\theta_{21}\theta_{22}.
\end{aligned}
$$

We now argue that the above allocation and transfer rules are strictly incentive compatible.

If player 1 of type θ_1 reports his type as $\hat{\theta}_1$ when player 2's type is θ_2, then player 1's payoff is

$$
\begin{aligned}
& p(\hat{\theta}_1,\theta_2)v_1(\theta_1,\theta_2) - z_1(\hat{\theta}_1,\theta_2) \\
&= p(\hat{\theta}_1,\theta_2)\big(v_1(\theta_1,\theta_2) - v_1(\hat{\theta}_1,\theta_2)\big) + \int_0^{\hat{\theta}_{11}} p((\tilde{\theta}_{11},\hat{\theta}_{12}),\theta_2)\frac{\partial v_1((\tilde{\theta}_{11},\hat{\theta}_{12}),\theta_2)}{\partial\theta_{11}}d\tilde{\theta}_{11} \\
&\quad + \int_0^{\hat{\theta}_{12}} p((\hat{\theta}_{11},\tilde{\theta}_{12}),\theta_2)\frac{\partial v_1((\hat{\theta}_{11},\tilde{\theta}_{12}),\theta_2)}{\partial\theta_{12}}d\tilde{\theta}_{12} - 2\hat{\theta}_{11}\hat{\theta}_{12} \\
&= \frac{1}{3}(\hat{\theta}_{11}+\theta_{21}+\hat{\theta}_{12}\theta_{22})\big((2+8\theta_{21}+9\theta_{22})(\theta_{11}-\hat{\theta}_{11}) + (1+4\theta_{21}+6\theta_{22})(\theta_{12}-\hat{\theta}_{12})\big) \\
&\quad + \frac{1}{3}\int_0^{\hat{\theta}_{11}} (\tilde{\theta}_{11}+\theta_{21}+\hat{\theta}_{12}\theta_{22})(2+8\theta_{21}+9\theta_{22})d\tilde{\theta}_{11} \\
&\quad + \frac{1}{3}\int_0^{\hat{\theta}_{12}} (\hat{\theta}_{11}+\theta_{21}+\tilde{\theta}_{12}\theta_{22})(1+4\theta_{21}+6\theta_{22})d\tilde{\theta}_{12} - 2\hat{\theta}_{11}\hat{\theta}_{12}
\end{aligned}
$$

After some calculation, we obtain

$$\begin{aligned}
E_{\theta_2}\frac{\partial}{\partial\hat{\theta}_{11}}\big(q(\hat{\theta}_1,\theta_2)v_1(\theta_1,\theta_2)-p_1(\hat{\theta}_1,\theta_2)\big) &= \frac{\partial}{\partial\hat{\theta}_{11}}E_{\theta_2}\big(q(\hat{\theta}_1,\theta_2)v_1(\theta_1,\theta_2)-p_1(\hat{\theta}_1,\theta_2)\big)\\
&= \frac{7}{2}(\theta_{11}-\hat{\theta}_{11})+2(\theta_{12}-\hat{\theta}_{12})\\
E_{\theta_2}\frac{\partial}{\partial\hat{\theta}_{12}}\big(q(\hat{\theta}_1,\theta_2)v_1(\theta_1,\theta_2)-p_1(\hat{\theta}_1,\theta_2)\big) &= \frac{\partial}{\partial\hat{\theta}_{12}}E_{\theta_2}\big(q(\hat{\theta}_1,\theta_2)v_1(\theta_1,\theta_2)-p_1(\hat{\theta}_1,\theta_2)\big)\\
&= 2(\theta_{11}-\hat{\theta}_{11})+\frac{7}{6}(\theta_{12}-\hat{\theta}_{12}).
\end{aligned}$$

It then follows that $\hat{\theta}_{11}=\theta_{11}$ and $\hat{\theta}_{12}=\theta_{12}$ is the unique maximizer of individual 1's expected payoff. A similar argument works for player 2. Hence, the allocation and transfer rules are strictly incentive compatible. As these rules are also continuous, it follows from Theorem 2b that they are continuously implementable when individuals' depth of reasoning is bounded. ◇

7 Conclusion

By imposing a bound on the agents' depth of reasoning, which we assume starts with truth-telling in simple direct mechanisms, we have presented results to show the permissiveness of mechanism design. In spite of requiring full implementation, incentive compatibility alone presents limitations to implementation with bounded depth of reasoning. Once small modeling mistakes are allowed, adding continuity to the mechanism, no other condition beyond incentive compatibility is required. The sufficiency counterparts of these results rely on the strict version of incentive compatibility. We have presented examples to showcase the applicability of the approach, which suggest new interesting directions for the theory of incentives without relying on the rational expectations assumption.

Appendix

This appendix provides the proofs of three lemmata used in the proof of Theorem 2b.

Proof of Lemma 1: For each $i \in I$, define $Z_i = \Delta\Theta \setminus Q_i^1(\hat{T})$. Let $d : \Delta\Theta \times \Delta\Theta \to \mathbb{R}_+$ be the Prohorov metric. Pick any $z_i \in Z_i$ and let $d(z_i, Q_i^1(\hat{T})) = \inf\{d(z_i, q_i^1) : q_i^1 \in Q_i^1(\hat{T})\}$. Since Z_i is open (recall that $Q_i^1(\hat{T})$ is closed by assumption), $d(z_i, Q_i^1(\hat{T})) > 0$. Let $B(z_i, \frac{d(z_i, Q_i^1(\hat{T}))}{4})$ be an open ball around z_i of radius $\frac{d(z_i, Q_i^1(\hat{T}))}{4}$. Note that $B(z_i, \frac{d(z_i, Q_i^1(\hat{T}))}{4}) \subset Z_i$. Now, $\left\{B(z_i, \frac{d(z_i, Q_i^1(\hat{T}))}{4})\right\}_{z_i \in Z_i}$ is an open cover of Z_i. Since Z_i is a metric space, it is paracompact. Therefore, the open cover $\left\{B(z_i, \frac{d(z_i, Q_i^1(\hat{T}))}{4})\right\}_{z_i \in Z_i}$ 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[11]

1. Each h_{z_i} is continuous.
2. Each $h_{z_i}(m_i) = 0$ if $m_i \in Z_i \setminus B(z_i, \frac{d(z_i, Q_i^1(\hat{T}))}{4})$.
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 $\rho_i(z_i) \in Q_i^1(\hat{T})$ be such that $d(z_i, \rho_i(z_i)) < \frac{5}{4} d(z_i, Q_i^1(\hat{T}))$.

[11] 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.

For each $i \in I$, define the correspondence $\omega_i : \Delta\Theta \to Q_i^1(\hat{T})$ as follows:

$$\omega_i(m_i) = \begin{cases} \{m_i\}, & \text{if } m_i \in Q_i^1(\hat{T}) \\ \{\rho_i(z_i) : z_i \in Z_i \text{ and } h_{z_i}(m_i) > 0\}, & \text{if } m_i \in Z_i. \end{cases}$$

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 $m_i \in \Delta\Theta$, define the probability distribution ξ_{m_i} over $Q_i^1(\hat{T})$ as follows:

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

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

Now, define $\mu : (\Delta\Theta)^I \to \Delta X$ as follows:

$$\mu(m) = \sum_{q^1 \in \times_{i \in I} \omega_i(m_i)} \times_{i \in I} \xi_{m_i}(q_i^1) \times \hat{\mu}(q^1).$$

Since $\mu(m) = \hat{\mu}(m), \forall m \in \times_{i \in I} Q_i^1(\hat{T})$, the mechanism μ is an extension of $\hat{\mu}$ to $(\Delta\Theta)^I$.

We now argue that μ is continuous. Let $(m^n)_{n \geq 1}$ be a sequence in $(\Delta\Theta)^I$ 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

$$\begin{aligned} I_1 &= \{i \in I : m_i \text{ is in } Z_i\} \\ I_2 &= \{i \in I : m_i \text{ is in the interior of } Q_i^1(\hat{T})\} \\ I_3 &= \{i \in I : m_i \text{ is on the boundary of } Q_i^1(\hat{T})\}. \end{aligned}$$

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 $Q_i^1(\hat{T})$, then there exists n_i^* such that $m_i^n \in Q_i^1(\hat{T})$ for all $n \geq n_i^*$.

Case 3. $i \in I_3$: In this case, m_i is on the boundary of $Q_i^1(\hat{T})$. Suppose the sequence $(m_i^n)_{n\geq 1}$ is such that it is infinitely often in Z_i – otherwise, the sequence $(m_i^n)_{n\geq 1}$ itself converges to m_i. Then consider its subsequence $(m_i^{n_l})_{n_l \geq 1}$ such that $m_i^{n_l} \in Z_i, \forall n_l \geq 1$. For each $m_i^{n_l}$, pick any $q_i^{1n_l} \in \omega_i(m_i^{n_l})$. Let $z_i^{n_l}$ be such that $\rho_i(z_i^{n_l}) = q_i^{1n_l}$ and $h_{z_i^{n_l}}(m_i^{n_l}) > 0$. We argue that the sequence $(q_i^{1n_l})_{n_l \geq 1}$ converges to m_i in the weak* topology.

To see this, 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 $d(m_i, q_i^{1n_l}) < \epsilon$ for all $n_l \geq n_i$. Note that

$$\begin{aligned} d(m_i, q_i^{1n_l}) &\leq d(m_i, m_i^{n_l}) + d(m_i^{n_l}, q_i^{1n_l}) \\ &\leq d(m_i, m_i^{n_l}) + d(m_i^{n_l}, z_i^{n_l}) + d(z_i^{n_l}, q_i^{1n_l}) \\ &< d(m_i, m_i^{n_l}) + d(m_i^{n_l}, z_i^{n_l}) + \frac{5}{4} d(z_i^{n_l}, Q_i^1(\hat{T})). \end{aligned}$$

Since $h_{z_i^{n_l}}(m_i^{n_l}) > 0$, we have $d(m_i^{n_l}, z_i^{n_l}) < \frac{d(z_i^{n_l}, Q_i^1(\hat{T}))}{4}$. Hence, $d(m_i, q_i^{1n_l}) < d(m_i, m_i^{n_l}) + \frac{6}{4} d(z_i^{n_l}, Q_i^1(\hat{T}))$.

Next,

$$d(z_i^{n_l}, Q_i^1(\hat{T})) \leq d(z_i^{n_l}, m_i) \leq d(z_i^{n_l}, m_i^{n_l}) + d(m_i^{n_l}, m_i) < \frac{d(z_i^{n_l}, Q_i^1(\hat{T}))}{4} + d(m_i^{n_l}, m_i).$$

Therefore, $\frac{3}{4} d(z_i^{n_l}, Q_i^1(\hat{T})) < d(m_i^{n_l}, m_i)$. As a result,

$$d(m_i, q_i^{1n_l}) < d(m_i, m_i^{n_l}) + \frac{6}{4} d(z_i^{n_l}, Q_i^1(\hat{T})) < 3 d(m_i, m_i^{n_l}) < \epsilon.$$

Hence, $(q_i^{1n_l})_{n_l \geq 1}$ 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

$$\mu(m^n)(A) = \sum_{q^1 \in \times_{i \in I} \omega_i(m_i^n)} \times_{i \in I} \xi_{m_i^n}(q_i^1) \times \hat{\mu}(q^1)(A).$$

Consider any $n \geq n^* = \max\{n_i^* : i \in I_1 \cup I_2\}$. Then m_i^n is in the interior of $Q_i^1(\hat{T}), \forall i \in I_2$. Hence,

$$\mu(m^n)(A) = \sum_{(q_i^1)_{i \in I_1 \cup I_3} \in \times_{i \in I_1 \cup I_3} \omega_i(m_i^n)} \times_{i \in I_1 \cup I_3} \xi_{m_i^n}(q_i^1) \times \hat{\mu}((q_i^1)_{i \in I_1 \cup I_3}, (m_i^n)_{i \in I_2})(A). \tag{7}$$

Pick any $(q_i^1)_{i \in I_3} \in \times_{i \in I_3} \omega_i(m_i^n)$, and define

$$Y^n((q_i^1)_{i \in I_3}) = \sum_{(q_i^1)_{i \in I_1} \in \times_{i \in I_1} \omega_i(m_i^n)} \times_{i \in I_1} \xi_{m_i^n}(q_i^1) \times \hat{\mu}((q_i^1)_{i \in I_1}, (q_i^1)_{i \in I_3}, (m_i^n)_{i \in I_2})(A).$$

Then it follows from (7) that

$$\mu(m^n)(A) = \sum_{(q_i^1)_{i \in I_3} \in \times_{i \in I_3} \omega_i(m_i^n)} \times_{i \in I_3} \xi_{m_i^n}(q_i^1) Y^n((q_i^1)_{i \in I_3}).$$

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

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

Take any $i \in I_1$ and $q_i^1 \in \{\rho_i(z_i) : z_i \in Z_i^*\}$. Since $m_i^n \in \mathcal{N}_i$, we have $\xi_{m_i^n}(q_i^1) = \sum_{z_i \in Z_i^* : \rho_i(z_i) = q_i^1} h_{z_i}(m_i^n)$. As each h_{z_i} is continuous,

$$\lim_{n\to\infty} \xi_{m_i^n}(q_i^1) = \sum_{z_i \in Z_i^* : \rho_i(z_i) = q_i^1} h_{z_i}(m_i) = \xi_{m_i}(q_i^1).$$

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

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

A similar argument shows that $\lim_{n\to\infty} Y^n((\tilde{q}_i^{1n})_{i\in I_3}) = \mu(m)(A)$. Therefore, μ is continuous. □

Proof of Lemma 2: We proceed by induction on k. The property is trivially satisfied when $k = 0$, as $\Sigma_i^0(\hat{t}_i) = \{q_i^1(\hat{t}_i)\}$. Suppose now that $k > 0$, and that the property holds for all $k' < k$. Let $m_i \in \Sigma_i^k(\hat{t}_i)$. By the induction hypothesis, individual i's conjecture has j of any type $\hat{t}_j \in \hat{T}_j$ and any cognitive state $c_j^{k'}$, where $k' < k$, report m_j such that $\omega_j(m_j) = \{q_j^1(\hat{t}_j)\}$. Since $\hat{T}$ is belief-closed, we must have

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

The strict incentive compatibility of $\hat{\mu}$ implies that $\omega_i(m_i) = \{q_i^1(\hat{t}_i)\}$, as desired. □

Proof of Lemma 3: We argue by induction that $Gr(S_i^k)$ is closed for all i and $k \geq 0$. Since $\Delta\Theta$ is compact, this implies that $S_i^k : T_i^* \times C_i^k \to \Delta\Theta$ is upper hemicontinuous for all i and $k \geq 0$. As C_i^k is compact, it is then straightforward to argue that Σ_i^k is upper hemicontinuous.

$Gr(S_i^0)$ is clearly closed for all i. Now suppose $Gr(S_i^{k'})$ is closed for all $k' \leq k-1$ and i. Pick any individual i and consider sequences $(t_i^n)_{n\geq 1}$, $(c_i^{kn})_{n\geq 1}$, and $(m_i^n)_{n\geq 1}$ such that $t_i^n \to t_i$, $c_i^{kn} \to c_i^k$, $m_i^n \to m_i$, and $m_i^n \in S_i^k(t_i^n, c_i^{kn}), \forall n$. Since $m_i^n \in S_i^k(t_i^n, c_i^{kn})$, there exists $\gamma^n \in \Delta(\Theta \times T_{-i}^* \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times (\Delta\Theta)^{I-1})$, 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 $\cup_{k'=0}^{k-1} C_{-i}^{k'}$ equals c_i^{kn}, *(c)* the marginal of γ^n on $T_{-i}^* \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times (\Delta\Theta)^{I-1}$ supports a subset of $\cup_{k'=0}^{k-1} \left(\times_{j\neq i} Gr(S_j^{k'})\right)$, and

$$m_i^n \in \arg\max_{m_i' \in \Delta\Theta} \int_{\Theta \times T_{-i}^* \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times (\Delta\Theta)^{I-1}} U_i(\mu(m_i', m_{-i}), \theta) d\gamma^n.$$

Since $\Theta \times T_{-i}^* \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times (\Delta\Theta)^{I-1}$ is a compact metric space, so is $\Delta(\Theta \times T_{-i}^* \times \cup_{k'=0}^{k-1} C_{-i}^{k} \times (\Delta\Theta)^{I-1})$. 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 $\text{marg}_{\Theta \times T_{-i}^*} \gamma = \pi_i(t_i)$. Similarly, $\text{marg}_{\cup_{k'=0}^{k-1} C_{-i}^{k'}} \gamma = c_i^k$.

By the induction hypothesis, $Gr(S_j^{k'})$ is closed for all $k' \leq k-1$ and $j \neq i$. Hence, $\Theta \times \cup_{k'=0}^{k-1} \left(\times_{j\neq i} Gr(S_j^{k'})\right)$ is closed. The fact that γ^{n_l} converges to γ in the weak* topology implies that

$$\gamma\left(\Theta \times \cup_{k'=0}^{k-1}\left(\times_{j\neq i} Gr(S_j^{k'})\right)\right) \geq \limsup_{n_l} \gamma^{n_l}\left(\Theta \times \cup_{k'=0}^{k-1}\left(\times_{j\neq i} Gr(S_j^{k'})\right)\right) = 1.$$

Therefore, the marginal of γ on $T_{-i}^* \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times (\Delta\Theta)^{I-1}$ supports a subset of $\cup_{k'=0}^{k-1} \left(\times_{j\neq i} Gr(S_j^{k'})\right)$.

Define

$$W_i(\hat{m}_i, \hat{\gamma}) = \int_{\Theta \times T_{-i}^* \times \cup_{k'=0}^{k-1} C_{-i}^{k} \times (\Delta\Theta)^{I-1}} U_i(\mu(\hat{m}_i, m_{-i}), \theta) d\hat{\gamma}.$$

We argue that W_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 $W_i(\hat{m}_i, \hat{\gamma}^n)$ converges to $W_i(\hat{m}_i, \hat{\gamma})$. That is, for every $\epsilon > 0$, there exists n_1 such that if $n \geq n_1$, then $|W_i(\hat{m}_i, \hat{\gamma}^n) - W_i(\hat{m}_i, \hat{\gamma})| < \frac{\epsilon}{2}$.

Since $U_i(\mu(m),\theta)$ is a continuous function over a compact metric space $(\Delta\Theta)^I \times \Theta$, 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 $(m_{-i},\theta) \in (\Delta\Theta)^{I-1} \times \Theta$. Therefore, for all $n \geq n_2$, we have

$$
\begin{aligned}
&|W_i(\hat{m}_i^n, \hat{\gamma}^n) - W_i(\hat{m}_i, \hat{\gamma}^n)| \\
&\leq \int_{\Theta \times T_{-i}^* \times \cup_{k'=0}^{k-1} C_{-i}^{k'} \times (\Delta\Theta)^{I-1}} |U_i(\mu(\hat{m}_i^n, m_{-i}),\theta) - U_i(\mu(\hat{m}_i, m_{-i}),\theta)| d\hat{\gamma}^n < \frac{\epsilon}{2}.
\end{aligned}
$$

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

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

Therefore, W_i is continuous. It follows from Berge's Maximum Theorem that $\arg\max_{\hat{m}_i \in \Delta\Theta} W_i(\hat{m}_i, \hat{\gamma})$ is upper hemicontinuous.

Now, returning to the subsequences $(m_i^{n_l})_{n_l \geq 1}$ and $(\gamma^{n_l})_{n_l \geq 1}$, we have $m_i^{n_l} \in \arg\max_{\hat{m}_i \in \Delta\Theta} W_i(\hat{m}_i, \gamma^{n_l})$. So we must have $m_i \in \arg\max_{\hat{m}_i \in \Delta\Theta} W_i(\hat{m}_i, \gamma)$. We thus conclude that $m_i \in S_i^k(t_i, c_i^k)$, and so $Gr(S_i^k)$ is closed. $\square$

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.

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.

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

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

Börgers, T., and T. Oh (2012). "Common Prior Type Spaces in Which Payoff Types and Belief Types are Independent." Mimeo, University of Michigan.

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.

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

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

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

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

Costa-Gomes, M. A. and V. P. Crawford (2006), "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review* 96, 1737-1768.

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

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

de Clippel, G., "Behavioral Implementation," *American Economic Review*, forthcoming.

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

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

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

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

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.

Lopomo, G., L. Rigotti, and C. Shannon (2009). "Uncertainty in Mechanism Design." Mimeo, University of Pittsburgh.

Matsushima, H. (1993), "Bayesian Monotonicity with Side Payments," *Journal of Economic Theory* 59, 107-121.

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.

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.

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

Saran, R. (2014). "Robust Implementation under Complete Information." Mimeo, Yale-NUS College.

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

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

Strzalecki, T. (2010), "Depth of Reasoning and Higher Order Beliefs," *mimeo*, Harvard.

Wang, J. T.-Y., M. Spezio, and C. F. Camerer (2010). "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games." *American Economic Review* 100, 984-1007.

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