**CMS-EMS**
**Center for Mathematical Studies in Economics**
**And Management Science**

Discussion Paper #1569

# Finite Depth of Reasoning and Equilibrium Play in Games with Incomplete Information

Willemien Kets
Northwestern University

February 7, 2014

*JEL Classification*: C700, C720, D800, D830

*Keywords*: Bounded rationality, higher-order beliefs, finite depth of reasoning, games with incomplete information, Bayesian-Nash equilibrium.

# Finite Depth of Reasoning and Equilibrium Play in Games with Incomplete Information

Willemien Kets*

February 7, 2014

**Abstract**

The standard framework for analyzing games with incomplete information models players as if they have an infinite depth of reasoning. This paper generalizes the type spaces of Harsanyi (1967–1968) so that players can have a finite depth of reasoning. We do this by restricting the set of events that a player of a finite depth can reason about. This allows us to extend the Bayesian-Nash equilibrium concept to environments with players with a finite depth of reasoning. We demonstrate that the standard approach of modeling beliefs with Harsanyi type spaces fails to capture the equilibrium behavior of players with a finite depth, at least in certain games. Consequently, the standard approach cannot be used to describe the equilibrium behavior of players with a finite depth in general. The same result can be shown to hold for rationalizability, showing that the results do not hinge on the specifics of the solution concept.

*JEL classification:* C700, C720, D800, D830

*Keywords:* Bounded rationality, higher-order beliefs, finite depth of reasoning, games with incomplete information, Bayesian-Nash equilibrium.

---

*Kellogg School of Management, Northwestern University. E-mail: w-kets@kellogg.northwestern.edu. Phone: +1-505-204 8012. This paper supersedes Kets (2009) and Kets (2010). I am grateful to Adam Brandenburger, Yossi Feinberg, and Matthew Jackson for their guidance and support, to Adam Brandenburger, Eddie Dekel, Alfredo Di Tillio, Jeff Ely, Amanda Friedenberg, Ben Golub, Joe Halpern, Aviad Heifetz, Philippe Jehiel, Rosemarie Nagel, Antonio Penta, Marcin Peski, Tomasz Sadzik, Dov Samet, Marciano Siniscalchi, Rani Spiegler, and Jonathan Weinstein for stimulating discussions, and to numerous seminar audiences for helpful comments. Part of this research was carried out during visits to Stanford University and the NYU Stern School of Business, and I thank these institutions for their hospitality. Financial support from the Air Force Office for Scientific Research under Grant FA9550-08-1-0389 is gratefully acknowledged.

# 1. Introduction

In games with incomplete information, it is important to not only consider players' beliefs about the state of nature, but also their beliefs about other players' beliefs. Consider, for example, a player who has to decide whether or not to invest in a project. The payoff associated with each choice depends on the economic fundamentals (i.e., the state of nature), as well as the actions of other investors. The player's optimal decision thus depends on her beliefs about the state of nature, i.e., on her first-order belief. Because the same is true for her opponents, the player's optimal action may also depend on her belief about her opponents' first-order belief, i.e., on her second-order belief. And because her opponents in turn condition their action on *their* beliefs about their opponents' beliefs about the state of nature, the player's optimal choice may also depend on her belief about her opponents' second-order beliefs (i.e., her third-order belief), and so on, ad infinitum.

Harsanyi (1967–1968) developed a tractable framework to analyze such games with incomplete information, and Harsanyi type spaces are widely used to study questions of economic interest. However, in the Harsanyi formalism, players are modeled as if they have an *infinite depth* of reasoning, that is, as if they can form beliefs about every possible higher-order event. Since it seems empirically plausible that players only have a *finite depth* of reasoning, it is important to understand the behavior of players with a finite depth of reasoning in games with incomplete information. In particular, an important question is whether the standard framework can be used to model the behavior of such players.

While there is an extensive literature on the behavior of players with a finite depth in games without payoff uncertainty, much less is understood about their behavior in games with incomplete information. This paper provides the first general framework that jointly models players' higher-order beliefs and their depth of reasoning, to analyze players' behavior in games with incomplete information.[1] The framework generalizes the standard Harsanyi framework to allow players to have a finite depth of reasoning, that is, to have beliefs only up to order four, say, or to think possible that an opponent has beliefs only up to order two or three.

The key innovation is that a player's depth is modeled by the set of events that a player can reason about, rather than by a simple number, as is common in the literature (footnote 1). For example, a player of depth 2, who can reason only about her opponents' first-order beliefs, can reason precisely about the events that can be described in terms of the first-order

[1]Play in games with complete information by players with a finite depth is studied by, e.g., Nagel (1995), Stahl and Wilson (1995), Ho et al. (1998), Costa-Gomes et al. (2001), Strzalecki (2009), and Alaoui and Penta (2013). Brocas et al. (2009), Crawford and Iriberri (2007), and Rogers et al. (2009) present behavioral models for games with incomplete information, but do not develop a model of beliefs independent of behavior. See Crawford et al. (2012) for a survey, and see Section 8 for further discussion.

beliefs of his opponents. As we discuss in Section 8, this extends the notion of a small world of Savage (1954) to a strategic context. We construct players' belief hierarchies, and show that each type generates a belief hierarchy of a well-defined depth (Theorem 5.2 and Corollary 5.3).

Our richer framework allows us to derive new strategic implications. Unlike the existing literature, we focus on (Bayesian-Nash) equilibrium for most of the paper, as this is the solution concept most commonly used in applications with incomplete information. Maintaining the assumption of equilibrium behavior also makes it possible to understand the effect of players' depth of reasoning in isolation. This allows us to study whether predictions obtained using the standard (equilibrium) framework remain valid when players have a finite depth of reasoning.[2]

A natural question is whether the equilibrium behavior of players with a finite depth of reasoning can be modeled using Harsanyi type spaces. One might hope, for instance, that for a given type space $\mathcal{T}$ in which players have a finite depth of reasoning, there is a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ that gives the same equilibria in every game as does $\mathcal{T}$. If that is the case, then we do not need to be concerned with the question whether or not real players have a finite or infinite depth if we are interested in equilibrium predictions: we can simply use the Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ to model equilibrium play.

The Harsanyi type spaces typically used in applied work are natural candidates for this purpose: in these type spaces, the higher-order beliefs of a type are determined uniquely by its beliefs up to some fixed, finite order.[3] Since strategies cannot depend on beliefs at arbitrarily high order by definition, one might hope that such Harsanyi type spaces can be used to model the behavior of players with a finite depth of reasoning.

A first question is whether for a given finite-depth type space, there is a Harsanyi type space with the property that for any game, for any equilibrium of the finite-depth type space, there is a corresponding equilibrium in the Harsanyi type space. Theorem 6.2 shows that the answer is yes, and that the Harsanyi type spaces for which this holds are precisely the Harsanyi type spaces in which it is common belief that the finite-order beliefs are as specified by the finite-depth type space. This includes the "finite-order" Harsanyi type spaces described above, but also much more complex ones; see, e.g., the proof of Lemma 6.1.

As for the converse, Theorem 6.4 proves a negative result: for every finite-depth type space, for every Harsanyi type space, there is a game with the property that there is an equilibrium

[2] Of course, it is an empirical question whether players (with a finite or infinite depth) actually follow equilibrium strategies, and, if so, under what conditions.

[3] In other words, there is some $k < \infty$ such that for each type in the Harsanyi type space, the higher-order beliefs it induces are commonly known conditional on its $k$th-order beliefs; see, e.g., Morris et al. (1995) and Qin and Yang (2013). This class of type spaces includes Harsanyi type spaces with finite type sets, or Harsanyi type spaces in which types represent payoffs and are drawn from a common prior.

for the Harsanyi type space such that there is no corresponding equilibrium for the finite-depth type space. This means that, at least in general games, Harsanyi type spaces are not suitable to model the equilibrium behavior of players with a finite depth of reasoning.

The intuition behind Theorem 6.4 is simple, but subtle. In every Harsanyi type space, strategies can depend on beliefs at higher order, and such strategies cannot be part of an equilibrium in a finite-depth space because players' expected utility cannot be calculated when strategies are too complex in that sense. The subtlety lies in understanding why even finite-order Harsanyi type spaces do not "work." After all, in such Harsanyi type spaces, a type's higher-order beliefs are determined completely by its belief up to some finite order $k$, so that strategies that depend on players' higher-order beliefs effectively depend on beliefs up to order (at most) $k$.

However, while the expected utility of such strategies for players of depth (at least) $k + 1$ is well-defined, it turns out that if types have finite depth $k + 1$ (or higher), then their beliefs cannot be captured by a Harsanyi type space in which beliefs are determined by players' beliefs up to order $k$. Recall that a type has depth $k + 1$ precisely when the set of events it can reason about are the events that can be described in terms of its opponents' $k$th-order beliefs. But in a Harsanyi type space in which beliefs are determined by the beliefs up to order $k$, this set coincides with the set of events that can be described in terms of the other players' $(k + 1)$th-order beliefs. This means that a type with such beliefs would have beliefs at least $k + 2$, a contradiction.

The richer framework for modeling players' depth is thus critical for understanding why Harsanyi type spaces cannot model players' equilibrium behavior when they have a finite depth of reasoning. In particular, if we were to define type spaces in the usual way and just add a parameter that specifies the depth of reasoning of the type, then we would not be able to see that the finite-order Harsanyi type spaces described above cannot capture the equilibrium behavior of types with a finite depth of reasoning.

One interpretation of our results is that to capture the behavior of players with a finite depth of reasoning, we need to consider a refinement of Bayesian-Nash equilibrium: rather than considering all equilibria, we should consider only the equilibria that depend on beliefs at sufficiently low order. We define such a refinement and show that this refinement indeed captures the equilibrium behavior of finite-depth type provided that we choose the Harsanyi type space appropriately.

Our results are not particular to equilibrium: we demonstrate that the same results hold for (interim correlated) rationalizability, a concept that is defined by the iterated deletion of dominated strategies (Dekel et al., 2007). The intuition behind these results is identical to that behind the equilibrium results, suggesting that the results do not depend on the specifics of the solution concept.

Together, our results imply that if we are interested in the equilibrium behavior of players with a finite depth of reasoning, then either we have to consider a refinement of Bayesian-Nash equilibrium, or we have to work with a finite-depth type space: if we consider the Bayesian-Nash equilibria for a Harsanyi type space, then we are bound to get too many equilibria in some games.

The next section illustrates our main results with some simple examples. The formal treatment starts in Section 3.

# 2. Examples

## 2.1. Harsanyi type spaces

As shown by Harsanyi (1967–1968), players' higher-order beliefs can be represented in a compact way using type spaces. In a *Harsanyi type space*, each player $i$ is endowed with a set $T_i$ of *types*, and associating with each type $t_i$ a *belief* (probability measure) $\beta_i(t_i)$ about $\theta$ and the other player's type. The function $\beta_i$ that maps each type for $i$ into a belief is assumed to be measurable. Each type generates a belief hierarchy, as the next example illustrates:

**Example 1.** The state of nature $\theta$ can be either high ($H$) or low ($L$), and each player $i = a, b$ has four types, labeled $t_i^1, \ldots, t_i^4$. The beliefs of each type are given in Figure 1.

| $\beta_a(t_a^1)$ | $H$ | $L$ |
|---|---|---|
| $t_b^1$ | 1 | 0 |
| $t_b^2$ | 0 | 0 |
| $t_b^3$ | 0 | 0 |
| $t_b^4$ | 0 | 0 |

| $\beta_a(t_a^2)$ | $H$ | $L$ |
|---|---|---|
| $t_b^1$ | 0 | 0 |
| $t_b^2$ | 0 | 0 |
| $t_b^3$ | 1 | 0 |
| $t_b^4$ | 0 | 0 |

| $\beta_b(t_b^1)$ | $H$ | $L$ |
|---|---|---|
| $t_a^1$ | 1 | 0 |
| $t_a^2$ | 0 | 0 |
| $t_a^3$ | 0 | 0 |
| $t_a^4$ | 0 | 0 |

| $\beta_b(t_b^2)$ | $H$ | $L$ |
|---|---|---|
| $t_a^1$ | 0 | 0 |
| $t_a^2$ | 0 | 0 |
| $t_a^3$ | 1 | 0 |
| $t_a^4$ | 0 | 0 |

| $\beta_a(t_a^3)$ | $H$ | $L$ |
|---|---|---|
| $t_b^1$ | 0 | 0 |
| $t_b^2$ | 0 | 1 |
| $t_b^3$ | 0 | 0 |
| $t_b^4$ | 0 | 0 |

| $\beta_a(t_a^4)$ | $H$ | $L$ |
|---|---|---|
| $t_b^1$ | 0 | 0 |
| $t_b^2$ | 0 | 0 |
| $t_b^3$ | 0 | 1 |
| $t_b^4$ | 0 | 0 |

| $\beta_b(t_b^3)$ | $H$ | $L$ |
|---|---|---|
| $t_a^1$ | 0 | 0 |
| $t_a^2$ | 0 | 1 |
| $t_a^3$ | 0 | 0 |
| $t_a^4$ | 0 | 0 |

| $\beta_b(t_b^4)$ | $H$ | $L$ |
|---|---|---|
| $t_a^1$ | 0 | 0 |
| $t_a^2$ | 0 | 0 |
| $t_a^3$ | 0 | 1 |
| $t_a^4$ | 0 | 0 |

Figure 1: A (Harsanyi) type space. The beliefs for types for Ann on the left, and those for Bob on the right; we write $x$ for the singleton $\{x\}$.

Types and their beliefs specify players' higher-order beliefs. For example, type $t_a^1$ for

Ann believes (with probability 1) that the state of nature is $H$, which specifies its *first-order belief* $\mu_a^1(t_a^1)$; of course, the other types $t_i$ also generate a first-order belief $\mu_i^1(t_i)$. Type $t_a^1$ also believes that Bob believes that $\theta = H$ (as it assigns probability 1 to type $t_b^1$, which believes that $\theta = H$). This specifies the *second-order belief* $\mu_a^2(t_a^1)$ induced by $t_a^1$, which is a probability measure on the set of states of nature and Bob's *first-order belief hierarchies* $H_b^1 := \{\mu_b^1(t_b) : t_b = t_b^1, \ldots, t_b^4\}$; again, the other types likewise generate a second-order belief. Type $t_a^1$ also induces a *third-order belief* $\mu_a^3(t_a^1)$ on the set of states of nature and Bob's *second-order belief hierarchies* $H_b^2 := \{(\mu_b^1(t_b), \mu_b^2(t_b)) : t_b = t_b^1, \ldots, t_b^4\}$: the type believes that Bob believes that Ann believes that $\theta = H$ (as $t_b^1$ assigns probability 1 to type $t_a^1$, which puts probability 1 on $\theta = H$).

We can continue this way, uncovering the $k$th-order belief $\mu_a^k(t_a^1)$ that $t_a^1$ generates for each $k$, with a $k$th-order belief being a probability measure on the set of states of nature and Bob's $(k-1)$*th-order belief hierarchies* $H_b^{k-1} := \{(\mu_b^1(t_b), \ldots, \mu_b^{k-1}(t_b)) : t_b = t_b^1, \ldots, t_b^4\}$. This gives the *belief hierarchy* $h_a(t_a^1) = (\mu_a^1(t_a^1), \mu_a^2(t_a^1), \ldots)$ induced by (or generated by) $t_a^1$. $\triangleleft$

We want to model belief hierarchies that potentially have a finite depth of reasoning. Let us start by considering the case of an infinite depth. We say that a belief hierarchy $h_a = (\mu_a^1, \mu_a^2, \ldots)$ for Ann has an *infinite depth (of reasoning)* if for each $k$, the $k$th-order belief $\mu_a^k$ can assign a probability to each event induced by Bob's $(k-1)$th-order belief hierarchies. This is the case only if the $\sigma$-algebra on which $\mu_a^k$ is defined can distinguish between the $(k-1)$th-order belief hierarchies for Bob that differ in their $(k-1)$th-order belief.

Hence, the belief hierarchy $(\mu_a^1, \mu_a^2, \ldots)$ has an infinite depth of reasoning if the first-order belief $\mu_a^1$ is a probability measure on $\mathcal{F}_\Theta$; the second-order belief $\mu_a^2$ is a probability measure on the $\sigma$-algebra $\mathcal{F}_\Theta \times \mathcal{F}_b^1$, where $\mathcal{F}_b^1$ is the $\sigma$-algebra on Bob's first-order belief hierarchies that includes the events expressible in his beliefs about $\Theta$; and so on.

More precisely, we define $\mathcal{F}_b^1$ to be the coarsest $\sigma$-algebra on Bob's first-order belief hierarchies that contain the sets

$$\{\mu_b^1 : E \in \Sigma(\mu_b^1), \mu_b^1(E) \geq p\},$$

with $\Sigma(\mu_b^1)$ the $\sigma$-algebra on which $\mu_b^1$ is defined, for any event $E \in \mathcal{F}_\Theta$ and every probability $p \in [0,1]$. For general $m$, assume that for each player $i$, the $\sigma$-algebra $\mathcal{F}_i^{m-2}$ on player $i$'s $(m-2)$th-order belief hierarchies has been defined. Then, let $\mathcal{F}_b^{m-1}$ be the coarsest $\sigma$-algebra on Bob's $(m-1)$th-order belief hierarchies that contains the sets

$$\big\{(\mu_b^1, \mu_b^2, \ldots, \mu_b^{m-1}) : E \in \Sigma(\mu_b^{m-1}), \mu_b^{m-1}(E) \geq p\big\} \tag{2.1}$$

for every probability $p \in [0,1]$ and every event $E$ in $\mathcal{F}_\Theta \times \mathcal{F}_a^{m-2}$ concerning $\theta$ and Ann's $(m-2)$th-order belief hierarchies; the $\sigma$-algebra $\mathcal{F}_a^{m-1}$ on Ann's $(m-1)$th-order belief hierarchies is defined similarly. Then, the belief hierarchy $(\mu_a^1, \mu_a^2, \ldots)$ has an *infinite depth of reasoning* if the first-order belief $\mu_a^1$ is a probability measure on $\mathcal{F}_\Theta$, and for $m > 1$, the $m$th-order belief $\mu_a^m$ is a probability measure on the $\sigma$-algebra $\mathcal{F}_\Theta \times \mathcal{F}_b^{m-1}$ on $\Theta$ and Bob's $(m-1)$th-order belief hierarchies.[4]

Going back to Example 1, the belief of each type $t_a$ for Ann is defined on the $\sigma$-algebra on Bob's type set that distinguishes each individual state $(\theta, t_b)$. This means in particular that for any $k$, the $k$th-order belief $\mu_a^k(t_a)$ induced by Ann's type can distinguish the $(k-1)$th-order belief hierarchies induced by Bob's types that differ in their $(k-1)$th-order beliefs.[5] So, the belief hierarchy induced by the type has an infinite depth of reasoning. The same is true, in fact, for any type in a Harsanyi type space, as should be expected; see Observation 1 below.

## 2.2. Finite depth of reasoning

We want to extend the Harsanyi approach to allow types to induce a belief hierarchy of finite depth, where, loosely speaking, a belief hierarchy has depth $k < \infty$ if it can form a belief only about the state of nature and the other players $(k-1)$th-order beliefs. To capture this, we restrict the set of events that a type can assign a probability to, that is, we let the type's belief be defined on a coarser $\sigma$-algebra. The next example demonstrates that Ann can form a belief only about Bob's first-order beliefs whenever the belief of a type for Ann about Bob's types is defined on a $\sigma$-algebra that distinguishes Bob's types only when they differ in their first-order belief (but not when their beliefs differ exclusively at higher order).

**Example 2.** Consider the type space in Figure 2. Each type $\tilde{t}_a$ for Ann is endowed with the $\sigma$-algebra $\Sigma_a(\tilde{t}_a)$ generated by the partition $\{\{\tilde{t}_b^1, \tilde{t}_b^2\}, \{\tilde{t}_b^3, \tilde{t}_b^4\}\}$, and likewise for the types for Bob.

Each type $\tilde{t}_a$ for Ann generates a first-order belief $\mu_a^1(\tilde{t}_a)$. Type $\tilde{t}_a^1$, for example, believes that the state of nature is $H$. Each type $\tilde{t}_a$ also induces a second-order belief $\mu_a^2(\tilde{t}_a)$. Type $\tilde{t}_a^1$, for example, assigns probability 1 to the event that Bob has type $\tilde{t}_b^1$ or $\tilde{t}_b^2$ (i.e., to $\{\tilde{t}_b^1, \tilde{t}_b^2\}$), and thus to the event that Bob believes that the state of nature is $H$ (since both $\tilde{t}_b^1$ and $\tilde{t}_b^2$ assign probability 1 to $H$). However, type $\tilde{t}_a^1$ cannot say whether or not Bob believes that Ann believes that $\theta = H$. The reason is that $\tilde{t}_b^1$ and $\tilde{t}_b^2$ differ in their beliefs about Ann's belief about nature, and $\tilde{t}_a^1$ cannot assign a probability to the individual types. The third-order belief $\mu_a^3(\tilde{t}_a^1)$ therefore cannot assign a probability to every event involving Bob's second-order

[4]Definition 1 below is of a different form, but it is equivalent to the current one, by Lemma A.2. Taking $\mathcal{F}_b^{m-1}$ to be the *coarsest* $\sigma$-algebra that contains the sets in (2.1) is standard.

[5]Of course, the formal result requires relating the $\sigma$-algebra on Bob's type set to the $\sigma$-algebra $\mathcal{F}_b^{k-1}$ on Bob's $(k-1)$th-order beliefs. The proof of Lemma 5.1 makes this connection. Also see Corollary 5.3.

| $\beta_a(\tilde{t}_a^1)$ | $H$ | $L$ |
|---|---|---|
| $\{\tilde{t}_b^1, \tilde{t}_b^2\}$ | 1 | 0 |
| $\{\tilde{t}_b^3, \tilde{t}_b^4\}$ | 0 | 0 |

| $\beta_a(\tilde{t}_a^2)$ | $H$ | $L$ |
|---|---|---|
| $\{\tilde{t}_b^1, \tilde{t}_b^2\}$ | 0 | 0 |
| $\{\tilde{t}_b^3, \tilde{t}_b^4\}$ | 1 | 0 |

| $\beta_b(\tilde{t}_b^1)$ | $H$ | $L$ |
|---|---|---|
| $\{\tilde{t}_a^1, \tilde{t}_a^2\}$ | 1 | 0 |
| $\{\tilde{t}_a^3, \tilde{t}_a^4\}$ | 0 | 0 |

| $\beta_b(\tilde{t}_b^2)$ | $H$ | $L$ |
|---|---|---|
| $\{\tilde{t}_a^1, \tilde{t}_a^2\}$ | 0 | 0 |
| $\{\tilde{t}_a^3, \tilde{t}_a^4\}$ | 1 | 0 |

| $\beta_a(\tilde{t}_a^3)$ | $H$ | $L$ |
|---|---|---|
| $\{\tilde{t}_b^1, \tilde{t}_b^2\}$ | 0 | 1 |
| $\{\tilde{t}_b^3, \tilde{t}_b^4\}$ | 0 | 0 |

| $\beta_a(\tilde{t}_a^4)$ | $H$ | $L$ |
|---|---|---|
| $\{\tilde{t}_b^1, \tilde{t}_b^2\}$ | 0 | 0 |
| $\{\tilde{t}_b^3, \tilde{t}_b^4\}$ | 0 | 1 |

| $\beta_b(\tilde{t}_b^3)$ | $H$ | $L$ |
|---|---|---|
| $\{\tilde{t}_a^1, \tilde{t}_a^2\}$ | 0 | 1 |
| $\{\tilde{t}_a^3, \tilde{t}_a^4\}$ | 0 | 0 |

| $\beta_b(\tilde{t}_b^4)$ | $H$ | $L$ |
|---|---|---|
| $\{\tilde{t}_a^1, \tilde{t}_a^2\}$ | 0 | 0 |
| $\{\tilde{t}_a^3, \tilde{t}_a^4\}$ | 0 | 1 |

Figure 2: A type space in which types have depth 2.

belief. $\triangleleft$

To model that players can have a finite depth of reasoning (and potentially different depths), we define a type space in which a player's belief can be defined on different $\sigma$-algebras. Thus, we endow Bob's type set $T_b$ with a collection $\mathcal{S}_b$ of $\sigma$-algebras, rather than a single one, as in Harsanyi type spaces. The belief $\beta_a(t_a)$ of a type $t_a$ for Ann about Bob's type is defined on a $\sigma$-algebra $\Sigma_a(t_a)$ in $\mathcal{S}_b$, and similarly with the player labels interchanged. The $\sigma$-algebra $\Sigma_a(t_a)$ specifies the events that $t_a$ can reason about: the type can assign a probability only to events in $\Sigma_a(t_a)$, but not to other events. In Example 2, type $\tilde{t}_a^1$ can assign a probability to the event that Bob has type $\tilde{t}_b^1$ or $\tilde{t}_b^2$ (and thus to the event that Bob believes that $\theta = H$), but not to the event that Bob has type $\tilde{t}_b^1$ (and therefore not to the event that Bob believes that Ann believes that $\theta = H$).

Types for Ann that have a different $\sigma$-algebra have a different depth of reasoning, as we will see; likewise for Bob. This means that players may be uncertain about the depth of reasoning of their opponent; see Example 7 below for an illustration.

We say that a belief hierarchy $h_a = (\mu_a^1, \mu_a^2, \ldots)$ has finite depth of reasoning $k < \infty$ if for any $m \leq k$, the $m$th-order belief $\mu_a^m$ can assign a probability to all events expressible in terms of Bob's $(m-1)$th-order belief hierarchies, as before, while for $m > k$, its $m$th-order belief can assign a probability only to those events regarding Bob's $(m-1)$th-order belief hierarchies *that are expressible in terms of his $(k-1)$th-order belief hierarchies.* That is, $h_a$ has *finite depth (of reasoning) $k < \infty$* if

- for $m \leq k$, the $m$th-order belief $\mu_a^m$ is defined on $\mathcal{F}_\Theta \times \mathcal{F}_b^{m-1}$ on $\Theta$ and Bob's $(m-1)$th-order belief hierarchies, as in (2.1); and

- for $m > k$, the $m$th-order belief $\mu_a^m$ is defined on the $\sigma$-algebra $\mathcal{F}_\Theta \times \mathcal{F}_{b,k-1}^{m-1}$, with $\mathcal{F}_{b,k-1}^{m-1}$ the coarsest $\sigma$-algebra that contains the events that are expressible in terms of Bob's

$(k-1)$th-order belief hierarchies, that is, the events

$$\left\{(\mu_b^1, \mu_b^2, \ldots, \mu_b^{m-1}) : E \in \Sigma(\mu_b^{k-1}), \mu_b^{k-1}(E) \geq p\right\}$$

for $E \in \mathcal{F}_\Theta \times \mathcal{F}_a^{k-2}$ and $p \in [0,1]$, and $\mathcal{F}_{b,k-1}^{m-1} \subsetneq \mathcal{F}_{b,k}^{m-1}$, where we define $\mathcal{F}_{b,m-1}^{m-1} := \mathcal{F}_b^{m-1}$.

(The condition that $\mathcal{F}_{b,k-1}^{m-1}$ is a strict subset of $\mathcal{F}_{b,k}^{m-1}$ ensures that the depth of a belief hierarchy is well-defined; see Definition 1 below.) With some abuse of terminology, we say that a type has depth $k$ if it generates a belief hierarchy of depth $k$.

In Example 2, the higher-order beliefs $\mu_a^k(\tilde{t}_a^1)$, $k \geq 2$, induced by $\tilde{t}_a^1$ can assign a probability only to events that are expressible in terms of the state of nature and Bob's first-order beliefs: the $\sigma$-algebra $\Sigma_a(\tilde{t}_a^1)$ separates the types for Bob if and only if they differ in their beliefs about the state of nature, but lumps them together otherwise. This means that any event in $\Sigma_a(\tilde{t}_a^1)$ can be described in terms of Bob's first-order beliefs. The same is true for the other types. It follows that every type has depth 2.

Of course, if we let the $\sigma$-algebras in $\mathcal{S}_a$ and $\mathcal{S}_b$ be arbitrary, then a type need not generate a belief hierarchy of a well-defined depth. Lemma 3.1 and Theorem 5.2 demonstrate, though, that if we relax the condition that belief maps be measurable (as required in Harsanyi type spaces) in an appropriate way, then each type generates a belief hierarchy of a well-defined depth.

## 2.3. Equilibrium

### *2.3.1. Definition*

What can we say about the equilibrium play of players with a finite depth of reasoning? It will be convenient to restrict attention to type spaces in which each type has the same depth of reasoning $k \leq \infty$. We refer to type spaces in which all types have the same finite depth $k$ as a *depth-$k$ (type) space*; recall that a type space in which every type has an infinite depth of reasoning is a Harsanyi type space. Then, a strategy profile $\sigma = (\sigma_a, \sigma_b)$ is an *equilibrium* if for each player $i = a, b$, the following hold:

- for each type $t_i$, the expected utility of $t_i$ of each action $s_i$ that $i$ might play is well-defined if $j \neq i$ follows the strategy $\sigma_j$; and
- for each type $t_i$, every action $s_i$ that is played with positive probability under $\sigma_i(t_i)$ is a best response to $\sigma_j$, $j \neq i$.

This definition of course coincides with the definition of Bayesian-Nash equilibrium when $k = \infty$. If the first condition holds, then we say that $\sigma_j$ is *comprehensible* for $t_j$. A sufficient condition for a strategy to be comprehensible for a type is that it is measurable (with respect to the $\sigma$-algebra of the type).

The next example illustrates that at least in some games, the equilibrium play of types with a finite depth and types with an infinite depth coincides:

**Example 3.** Consider the game in Figure 3, where Ann is the row player, and Bob is the column player. Suppose that players' beliefs are given by the type space in Example 2, in which the depth of reasoning of each type equals 2.

| $\theta = H$ | $s_b^1$ | $s_b^2$ |
|---|---|---|
| $s_a^1$ | 1,1 | 1,0 |
| $s_a^2$ | 0,1 | 0,0 |

| $\theta = L$ | $s_b^1$ | $s_b^2$ |
|---|---|---|
| $s_a^1$ | 0,0 | 0,1 |
| $s_a^2$ | 1,0 | 1,1 |

Figure 3: A game with dominant actions.

It is easy to see that this game has a unique equilibrium in which types $\tilde{t}_a^1$ and $\tilde{t}_a^2$ play $s_a^1$ (with probability 1), as $s_a^1$ is a dominant action for types that believe that $\theta = H$, and types $\tilde{t}_a^3$ and $\tilde{t}_a^4$ play $s_a^2$, as $s_a^2$ is dominant for types that believe that $\theta = L$; and likewise for Bob.

Now suppose players' beliefs are given by the type space in Example 1. This type space generates the same second-order belief hierarchies as the type space in Example 2; for example, types $t_a^1$ and $\tilde{t}_a^1$ (in Example 1 and 2, respectively) both believe that $\theta = H$ and that Bob believes that $\theta = H$. However, each type in Example 1 has an infinite depth of reasoning. Again, there is a unique equilibrium in which types $t_a^1$ and $t_a^2$ play $s_a^1$ (with probability 1), and types $t_a^3$ and $t_a^4$ play $s_a^2$, and similarly for Bob's types. $\triangleleft$

Example 3 shows that for some depth-2 spaces and some games, there exist Harsanyi type spaces with the same set of equilibria. We want to know whether we can always find a Harsanyi type space that "mimics" the equilibrium predictions of a given finite-depth type space.

### *2.3.2. Strategic equivalence*

To answer that question, we need some more definitions. A *game* $\mathcal{G}$ specifies a set of actions for each player, as well as the players' payoffs for each action profile and state of nature. A *model* is a pair $(\mathcal{G}, \mathcal{T})$, with $\mathcal{G}$ a game, and $\mathcal{T}$ a type space. Roughly speaking, a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ is *strategically equivalent* to a depth-$k$ space $\mathcal{T}^k$ if for any game $\mathcal{G}$, the set of equilibria of the Harsanyi model $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$ "coincides" with the set of equilibria of the finite-depth model $(\mathcal{G}, \mathcal{T}^k)$, in the following sense: if $\mathcal{T}^{\mathcal{H}}$ and $\mathcal{T}^k$ have the same type sets (i.e., $T_i^H = T_i^k$ for each player $i$), then the spaces are strategically equivalent if

(1) any equilibrium of the depth-$k$ model $(\mathcal{G}, \mathcal{T}^k)$ is an equilibrium of the Harsanyi model $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$; and

(2) any equilibrium of the Harsanyi model $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$ is an equilibrium of the depth-$k$ model $(\mathcal{G}, \mathcal{T}^k)$.

If the type sets in $\mathcal{T}^{\mathcal{H}}$ and $\mathcal{T}^k$ are allowed to differ, then the definition is more involved; we leave the details for Section 6.2.

Conditions (1) and (2) are satisfied for the game in Figure 3 by the type spaces in Example 3. The question is whether for a given depth-$k$ type space, there is a Harsanyi type space that satisfies (1) and (2) for *every* game.

### *2.3.3. Harsanyi extensions*

We first ask whether, for a given depth-$k$ type space $\mathcal{T}^k$, condition (1) holds (for all games $\mathcal{G}$), that is, of there is a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ such that for any game $\mathcal{G}$, for any equilibrium $\sigma^k$ of the game when beliefs are given by $\mathcal{T}^k$, there is a corresponding equilibrium $\sigma$ of the game when beliefs are given by $\mathcal{T}^{\mathcal{H}}$. Not surprisingly, this does not hold for all Harsanyi type spaces:

**Example 4.** Consider the game in Figure 4, and consider the following type space. Each player $i = a, b$ has two types, labeled $t_i^1, t_i^2$. Each type $t_i$ is endowed with the trivial $\sigma$-algebra $\{\{t_j^1, t_j^2\}, \emptyset\}$ on the type set of the other player $j$. Type $t_a^1$ assigns probability $\frac{9}{10}$ to the event that $\theta = H$ (and that Bob has a type in $\{t_b^1, t_b^2\}$), and the complementary probability to the event that $\theta = L$. Type $t_a^2$ assigns probability $\frac{4}{5}$ to the event that $\theta = H$, and the remaining probability to the event that $\theta = L$. The beliefs for Bob's types are defined similarly. Since the $\sigma$-algebra on which the types' beliefs are defined lumps together types that differ in their first-order beliefs, each type has depth 1.

| | $s_b^1$ | $s_b^2$ |
|---|---|---|
| $s_a^1$ | 1,1 | -2,0 |
| $s_a^2$ | 0,-2 | 0,0 |

$\theta = H$

| | $s_b^1$ | $s_b^2$ |
|---|---|---|
| $s_a^1$ | -2,-2 | -2,0 |
| $s_a^2$ | 0,-2 | 0,0 |

$\theta = L$

Figure 4: A risky coordination game.

It is easy to see that this model has an equilibrium $\sigma$ in which type $t_i$ plays the risky action $s_i^1$, for $i = a, b$. Clearly, if beliefs are given instead by a (Harsanyi) type space in which all types have very different first-order beliefs than $t_i^1$ and $t_i^2$, $i = a, b$, – for example, assigning high probability to $\theta = L$ –, then this may no longer be an equilibrium.

However, even if beliefs are given by a (Harsanyi) type space which contains types that have the same beliefs as $t_i^1$ and $t_i^2$, there may not be an equilibrium in which types that assign a high probability to $\theta = H$ play the risky action $s_i^1$. To wit, we can construct a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ that for each player $i = a, b$, contains types $t_i^{\mathcal{H},1}$ and $t_i^{\mathcal{H},2}$ that generate the same first-order beliefs as $t_i^1$ and $t_i^2$, respectively, yet in *every* equilibrium of the game, types $t_i^{\mathcal{H},1}$ and $t_i^{\mathcal{H},2}$ play the safe action $s_i^2$. We can do this by including a type $t^*$ in $\mathcal{T}^{\mathcal{H}}$ that believes that $\theta = L$, so that the safe action is strictly dominant for $t^*$, and then choose $t_i^{\mathcal{H},1}$ and $t_i^{\mathcal{H},2}$ such that these types have the same first-order beliefs as $t_i^1$ and $t_i^2$, respectively, and that believe that the other player believes...that the other player has type $t^*$ (cf. Rubinstein, 1989). $\triangleleft$

Example 4 suggest that a necessary condition for (1) to hold is that $\mathcal{T}^{\mathcal{H}}$ does not include types that have different $k$th-order beliefs than the types in $\mathcal{T}^k$. More precisely, say that a Harsanyi type space is a *Harsanyi extension* of a depth-$k$ type space $\mathcal{T}^k$ if it generates exactly the same $k$th-order belief hierarchies as $\mathcal{T}^k$, and this is common belief. For example, the type space in Example 1 is a Harsanyi extension of the depth-2 space in Example 2.[6] So, the example suggests that a necessary condition for (1) to hold is that the Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ is a Harsanyi extension of the finite-depth space $\mathcal{T}^k$.

Theorem 6.2 shows that this is true, and that the converse also holds: (1) holds for any Harsanyi extension of the finite-depth space. To see this, note that the possibilities for profitable deviations do not change when we "increase" the depth of a type, while the condition that strategies be comprehensible in equilibrium becomes easier to satisfy.

Since there is a Harsanyi extension of a large class of depth-$k$ spaces (Lemma 6.1), it follows that for any depth-$k$ space $\mathcal{T}^k$ in that class, there is a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ such that for every game $\mathcal{G}$, for every equilibrium of the depth-$k$ model $(\mathcal{G}, \mathcal{T}^k)$, there is a a corresponding equilibrium of the Harsanyi model $(\mathcal{G}, \mathcal{T}^k)$. However, the converse of this result is false, as we discuss next.

### *2.3.4. A negative result*

The next example illustrates why the converse may be false:

**Example 5.** Consider the game in Figure 5. When the action set is restricted to $\{s_i^1, s_i^2\}$ for each player $i$, then players play a game with dominant actions, a coordination game, or an anti-coordination game, depending on the state. When players coordinate on $(s_a^2, s_b^3)$, their payoffs are independent of the state.

[6] Of course, type spaces have many different Harsanyi extensions. For example, any variant of the type space in Example 1 in which $t_a^1$ puts probability $p$ on $(H, t_b^1)$ and $1-p$ on $(H, t_b^2)$ for some $p \in [0,1]$ is a Harsanyi extension of the type space in Example 2. The proof of Lemma 6.1 constructs the "canonical" Harsanyi extension for a given depth-$k$ space that contains all other Harsanyi extensions.

| | $s_b^1$ | $s_b^2$ | $s_b^3$ |
|---|---|---|---|
| $s_a^1$ | 1, 1 | 1, 0 | 1, 0 |
| $s_a^2$ | 0, 1 | 0, 0 | 0, 0 |
| $s_a^3$ | 0, 1 | 0, 0 | 1, 1 |

$\theta = \theta_1$

| | $s_b^1$ | $s_b^2$ | $s_b^3$ |
|---|---|---|---|
| $s_a^1$ | 0, 0 | 0, 1 | 0, 0 |
| $s_a^2$ | 1, 0 | 1, 1 | 1, 0 |
| $s_a^3$ | 0, 0 | 0, 1 | 1, 1 |

$\theta = \theta_2$

| | $s_b^1$ | $s_b^2$ | $s_b^3$ |
|---|---|---|---|
| $s_a^1$ | 1, 1 | 0, 0 | 0, 0 |
| $s_a^2$ | 0, 0 | 1, 1 | 0, 0 |
| $s_a^3$ | 0, 0 | 0, 0 | 1, 1 |

$\theta = \theta_3$

| | $s_b^1$ | $s_b^2$ | $s_b^3$ |
|---|---|---|---|
| $s_a^1$ | 0, 0 | 1, 1 | 0, 0 |
| $s_a^2$ | 1, 1 | 0, 0 | 0, 0 |
| $s_a^3$ | 0, 0 | 0, 0 | 1, 1 |

$\theta = \theta_4$

Figure 5: A game for which some equilibria depend on second-order beliefs (given the type space).

Consider the type space $\mathcal{T}^2$, defined as follows. Each player $i = a, b$ has eight types, labeled $t_i^1, \ldots, t_i^8$, and each type $t_i$ for player $i$ is endowed with the $\sigma$-algebra $\mathcal{F}_j$ on $j$'s type set that is generated by the pairs $\{t_j^1, t_j^2\}$, $\{t_j^3, t_j^4\}$, $\{t_j^5, t_j^6\}$, and $\{t_j^7, t_j^8\}$, where $j \neq i$. The beliefs for player $i = a, b$ are given by:

$$\begin{aligned} \beta_i(t_i^1)(\theta_1, \{t_j^1, t_j^2\}) =& 1, & \beta_i(t_i^2)(\theta_1, \{t_j^3, t_j^4\}) =& 1; \\ \beta_i(t_i^3)(\theta_2, \{t_j^1, t_j^2\}) =& 1, & \beta_i(t_i^4)(\theta_2, \{t_j^3, t_j^4\}) =& 1; \\ \beta_i(t_i^5)(\theta_3, \{t_j^1, t_j^2\}) =& 1, & \beta_i(t_i^6)(\theta_3, \{t_j^3, t_j^4\}) =& 1; \\ \beta_i(t_i^7)(\theta_4, \{t_j^5, t_j^6\}) =& 1, & \beta_i(t_i^8)(\theta_4, \{t_j^7, t_j^8\}) =& 1, \end{aligned}$$

where $j \neq i$. It is straightforward to verify that the $\sigma$-algebra $\mathcal{F}_j$ separates the types for $j$ if they differ in their first-order belief, but not if they differ in their beliefs at higher order, so that each type $t_i$ has depth 2.

Clearly, the strategy profile $\sigma$ in which each type for player $i$ plays $s_i^3$ (with probability 1) is an equilibrium. Does this model have another equilibrium? For types $t_a^1$ and $t_a^2$ for Ann, it is a best response to play $s_a^1$, given their beliefs. Likewise, it is a best response for $t_a^3$ and $t_a^4$ to play $s_a^2$. In that case, the unique best response for type $t_b^5$ for Bob is to play $s_b^1$, while the unique best response for $t_b^6$ is to play $s_b^2$.

We cannot identify a best response for type $t_a^7$ to such a strategy, however. The problem is that we cannot calculate the expected payoff of $t_a^7$ to each of Ann's actions, i.e., Bob's strategy is not comprehensible. Similarly, we cannot calculate the expected payoff to $t_b^7$. And, of course, if we cannot determine the optimal behavior of $t_a^7$ and $t_b^7$, then it is unclear what the optimal play for $t_a^8$ and $t_b^8$ is. Hence, there is no equilibrium in which for some player $i$, types $t_i^1$ and $t_i^2$ play $s_i^1$, and types $t_i^3$ and $t_i^4$ play $s_i^2$. $\triangleleft$

The problem here is that there is a tension between the requirement that players choose a best response and the requirement that players' expected utility be well-defined. This tension does not arise in Harsanyi type spaces:

**Example 5 (cont.).** Refer back to the game in Figure 5, but now suppose that players' beliefs are given by the Harsanyi type space $\mathcal{T}^{\mathcal{H}}$, defined as follows. Again, each player $i = a, b$ has eight types, labeled $t_i^{\mathcal{H},1}, \ldots, t_i^{\mathcal{H},8}$, and each type $t_i^{\mathcal{H}}$ for player $i$ is endowed with the power set on $j$'s type set, where $j \neq i$. The beliefs for player $i$ are given by:

$$\begin{aligned} \beta_i^{\mathcal{H}}(t_i^{\mathcal{H},1})(\theta_1, t_j^{\mathcal{H},1}) &= 1, & \beta_i^{\mathcal{H}}(t_i^{\mathcal{H},2})(\theta_1, t_j^{\mathcal{H},3}) &= 1; \\ \beta_i^{\mathcal{H}}(t_i^{\mathcal{H},3})(\theta_2, t_j^{\mathcal{H},2}) &= 1, & \beta_i^{\mathcal{H}}(t_i^{\mathcal{H},4})(\theta_2, t_j^{\mathcal{H},4}) &= 1; \\ \beta_i^{\mathcal{H}}(t_i^{\mathcal{H},5})(\theta_3, t_j^{\mathcal{H},2}) &= 1, & \beta_i^{\mathcal{H}}(t_i^{\mathcal{H},6})(\theta_3, t_j^{\mathcal{H},3}) &= 1; \\ \beta_i^{\mathcal{H}}(t_i^{\mathcal{H},7})(\theta_4, t_j^{\mathcal{H},5}) &= 1, & \beta_i^{\mathcal{H}}(t_i^{\mathcal{H},8})(\theta_4, t_j^{\mathcal{H},7}) &= 1, \end{aligned}$$

where $j \neq i$.

It is easy to see that type $t_i^{\mathcal{H},m}$ generates the same second-order belief hierarchy as type $t_i^m$ in the original depth-2 space $\mathcal{T}^2$. Indeed, this type space is a Harsanyi extension of $\mathcal{T}^2$. Clearly, the strategy profile in which every type for player $i$ plays $s_i^3$ is still an equilibrium of this model.

But now there is another equilibrium $\sigma$ in which types $t_a^{\mathcal{H},1}$ and $t_a^{\mathcal{H},2}$ play $s_a^1$, and types $t_a^{\mathcal{H},3}$ and $t_a^{\mathcal{H},3}$ play $s_a^2$. In this case, the unique best responses for types $t_b^{\mathcal{H},5}$ and $t_b^{\mathcal{H},6}$ are $s_b^1$ and $s_b^2$, respectively. Given this, the unique best response for $t_a^{\mathcal{H},7}$ is to play $s_a^2$, and the unique best response for type $t_b^{\mathcal{H},8}$ is $s_b^1$; the best responses for the other types when $t_b^{\mathcal{H},1}$ and $t_b^{\mathcal{H},2}$ play $s_b^1$ and types $t_b^{\mathcal{H},3}$ and $t_b^{\mathcal{H},3}$ play $s_b^2$ can be determined likewise. $\triangleleft$

In Example 5, the strategy profile $\sigma$ is an equilibrium of the Harsanyi model, but not of the depth-2 model.

Theorem 6.4 shows that this holds generally: for every depth-$k$ type space $\mathcal{T}^k$, and for every Harsanyi extension $\mathcal{T}^{\mathcal{H}}$ of $\mathcal{T}^k$, there is a game $\mathcal{G}$ and an equilibrium of the Harsanyi model ($\mathcal{G}$, $\mathcal{T}^{\mathcal{H}}$) such that there is no corresponding equilibrium in the depth-$k$ model ($\mathcal{G}$, $\mathcal{T}^k$).

The critical insight behind Theorem 6.4 is that a type $t_i$ (say, type $t_a^7$ in Example 5) has depth 2 only if its opponent has two types $t_j, t_j'$ (say, $t_b^5$ and $t_b^6$) that have the same first-order belief, but differ in their second-order belief; otherwise, $t_i$ has depth at least 3.[7] In that case, type $t_i$ cannot distinguish $t_j$ and $t_j'$. However, it may be optimal for $t_j$ and $t_j'$ to play differently, as in the game in Figure 5. But if that is the case, the expected utility for $t_i$ need not be well-defined.

Note that there are games $\mathcal{G}$ such that a depth-$k$ model ($\mathcal{G}$, $\mathcal{T}^k$) does not have an equilibrium, even if for any Harsanyi extension $\mathcal{T}^{\mathcal{H}}$ of $\mathcal{T}^k$, the model ($\mathcal{G}$, $\mathcal{T}^{\mathcal{H}}$) has an equilibrium; a variant of the game in Example 5 in which the action set of each player $i = a, b$ is restricted

[7]Note that if no such types $t_j, t_j'$ exist, then $t_i$ can form beliefs about the second-order beliefs of player $j$; also see Corollary 5.3.

to $\{s_i^1, s_i^2\}$ is an example. While it may be desirable to have a prediction for a large class of games, we do not view the fact that certain (finite) models do not have an equilibrium as an anomaly: it simply reflects the fact that some equilibrium strategies are too complicated for players to reason about if they have limited cognitive resources. Indeed, in such a case, it may be more reasonable to assume nonequilibrium behavior; see, e.g., Brocas et al. (2009), Crawford and Iriberri (2007), Strzalecki (2009), and Heifetz and Kets (2013).[8]

### 2.4. Outline

The remainder of this paper is organized as follows. We construct players' belief hierarchies in Section 3. Section 4 introduces the type spaces we consider, and Section 5 shows that each type generates a belief hierarchy with a well-defined depth of reasoning. Section 6 presents the main results. Section 7 treats a number of conceptual and technical issues, and Section 8 discusses the related literature. Most proofs are relegated to the appendices.

## 3. Belief hierarchies

We now begin the formal treatment. In this section, we provide an explicit model of players' higher-order beliefs about the state of nature, by constructing their belief hierarchies. We do so in a way that every belief hierarchy has a well-defined depth of reasoning. Section 4 provides an implicit description of these beliefs, by generalizing the familiar Harsanyi representation.

### 3.1. Preliminaries

We start with some preliminaries. For a set $X$ and $\sigma$-algebra $\mathcal{F}$ on $X$, we write $\Delta(X, \mathcal{F})$ for the set of probability measures on $\mathcal{F}$ (i.e., on the measurable space $(X, \mathcal{F})$), and we endow $\Delta(X, \mathcal{F})$ with the $\sigma$-algebra $\mathcal{F}_{\Delta(X,\mathcal{F})}$ generated by the sets

$$\big\{\mu \in \Delta(X, \mathcal{F}) : \mu(E) \geq p\big\} : \qquad E \in \mathcal{F}, p \in [0, 1].$$

This $\sigma$-algebra naturally separates beliefs (probability measures) according to the probability they assign to events; this makes it possible to talk about "beliefs about beliefs," and so on (Heifetz and Samet, 1998). Moreover, this $\sigma$-algebra coincides with the Borel $\sigma$-algebra in the

[8] However, Theorem 6.4 does not follow directly from the fact that some finite-depth models do not have an equilibrium even if some models based on a Harsanyi extension of the finite-depth space do have equilibria. For this, one would have to show that for *every* depth-$k$ type space, with arbitrary type sets, there is a game for which the depth-$k$ model does not have an equilibrium, while the corresponding Harsanyi model does have an equilibrium for every Harsanyi extension of the depth-$k$ space.

common case that $\Delta(X, \mathcal{F})$ is endowed with the weak topology, $X$ is metrizable, and $\mathcal{F}$ is the Borel $\sigma$-algebra on $X$.

As is standard, the product of measurable spaces is endowed with the product $\sigma$-algebra, and a subset $Y$ of a space $X$, endowed with a $\sigma$-algebra $\mathcal{F}_X$, has the relative $\sigma$-algebra, denoted by $\mathcal{F}_Y$. If $\mu$ is a probability measure on a product space $X \times Y$, then its marginal on $X$ is denoted by $\operatorname{marg}_X \mu$.

For any family of spaces $\{X_z : z \in Z\}$, with $X_z$ endowed with the $\sigma$-algebra $\mathcal{F}_z$, $z \in Z$, the union $X := \bigcup_{z \in Z} X_z$ is endowed with the $\sigma$-algebra $\mathcal{F}$ that contains precisely the subsets $E \subseteq X$ such that $E \cap X_z \in \mathcal{F}_z$ for all $z \in Z$. That is, $(X, \mathcal{F})$ is the sum of the measurable spaces $(X_z, \mathcal{F}_z)$, $z \in Z$.[9] In particular, if $\mathscr{S}$ is a collection of $\sigma$-algebras on a space $Y$, then the space $\Delta(Y, \mathscr{S}) := \bigcup_{\mathcal{Q} \in \mathscr{S}} \Delta(Y, \mathcal{Q})$ is endowed with the $\sigma$-algebra generated by sets of the form

$$\left\{\mu \in \Delta(Y, \mathscr{S}) : \Sigma(\mu) = \mathcal{Q}, \mu(E) \geq p\right\} : \qquad \mathcal{Q} \in \mathscr{S}, E \in \mathcal{Q}, p \in [0, 1],$$

where $\Sigma(\mu)$ is the $\sigma$-algebra on which the belief $\mu$ is defined.

## 3.2. Construction

There is a set $N$ of players, who are uncertain about the *state of nature* $\theta \in \Theta$. The set $\Theta$ of states of nature is endowed with a $\sigma$-algebra $\mathcal{F}_\Theta$, and is assumed to contain at least two elements. For simplicity, we focus on the case of two players for much of the paper and write $N = \{a, b\}$; the results generalize to the case of three or more players with minor changes, see the online appendix. Throughout this paper, if we fix a player $i$, then the player other than $i$ is denoted by $j$, i.e., $j \neq i$.

Players form beliefs about $\theta$, about their opponent's beliefs about $\theta$, and so on. We define spaces of belief hierarchies to model players' higher-order beliefs, building on the construction of Mertens and Zamir (1985) for the Harsanyi case. We define the belief hierarchies in such a way that each belief hierarchy has a well-defined depth of reasoning; see Section 3.3.

Generalizing Definition 2.1 of Mertens and Zamir, we define a *space of belief hierarchies* to be a sequence $\boldsymbol{C} = (\boldsymbol{C}^1, \boldsymbol{C}^2, \ldots)$, with $\boldsymbol{C}^m = \prod_{i \in N} C_i^m$ for all $m$, that satisfies the following conditions:

$(i)$ For each player $i \in N$, $C_i^1 \subseteq \Delta(\Theta, \mathcal{F}_\Theta)$, and for $m = 2, 3, \ldots$,

$$C_i^m \subseteq C_i^{m-1} \times \Delta(\Theta \times C_j^{m-1}, \mathscr{S}_i^m(\boldsymbol{C}^{m-1})),$$

where the collection $\mathscr{S}_i^m(\boldsymbol{C})$ of $\sigma$-algebras is defined below;

[9] We implicitly assume here that the spaces $X_z$ are disjoint. This is without loss of generality: we can replace any space $X_z$ with an isomorphic copy if needed.

$(ii)$ For each player $i \in N$, $m = 1, 2, \dots$, and $(\mu_i^1, \dots, \mu_i^m) \in C_i^m$, we have $\operatorname{marg}_\Theta \mu_i^2 = \mu_i^1$, and $\operatorname{marg}_{\Theta \times C_j^{m-2}} \mu_i^m = \mu_i^{m-1}$ for $m > 2$.

$(iii)$ For each player $i \in N$ and $m = 1, 2, \dots$, the projection of $C_i^{m+1}$ into $C_i^{m-1} \times \Delta(\Theta \times C_j^{m-1}, \mathscr{S}_i^m(\boldsymbol{C}^{m-1}))$ equals $C_i^m$.

An example of such a space of belief hierarchies is the universal type space of Mertens and Zamir (1985); see Example 6 below.

Condition $(i)$ says that an *mth-order belief hierarchy* $(\mu_i^1, \dots, \mu_i^m) \in C_i^m$ consists of an $(m-1)$th-order belief hierarchy $(\mu_i^1, \dots, \mu_i^{m-1})$ and a belief $\mu_i^m$ about the state of nature and the other player's $(m-1)$th-order belief hierarchy. The belief $\mu_i^m$ is called the *mth-order belief* (induced by the hierarchy). We return to condition $(i)$ below. Condition $(ii)$ is a standard coherency condition that says that beliefs at different orders cannot contradict each other (cf. Mertens and Zamir, 1985; Brandenburger and Dekel, 1993). Condition $(iii)$ says that every $m$th-order belief hierarchy can be extended to an $(m+1)$th-order belief hierarchy. It is straightforward to show that this condition can be satisfied whenever $C_i^{m-1}$ is nonempty for every player $i$.

Thus, we obtain a sequence $C_i^1, C_i^2, \dots$ of spaces of finite-order belief hierarchies for each player $i$. A *belief hierarchy* for player $i$ (in $\boldsymbol{C}$) is a sequence $(\mu_i^1, \mu_i^2, \dots)$ of $m$th-order beliefs $\mu_i^m$, $m \geq 1$, such that for every $\ell$, we have $(\mu_i^1, \dots, \mu_i^\ell) \in C_i^\ell$. Thus, the set of belief hierarchies for player $i$ (in $\boldsymbol{C}$) is

$$H_i(\boldsymbol{C}) := \{(\mu_i^1, \mu_i^2, \dots) : \text{for all } \ell,\ (\mu_i^1, \dots, \mu_i^\ell) \in C_i^\ell\}.$$

Returning to Condition $(i)$, we note that it generalizes a similar condition of Mertens and Zamir (1985, Definition 2.1) by allowing the beliefs of player $i$ at a given order $m$ to be defined on different $\sigma$-algebras. The collection $\mathscr{S}_i^m(\boldsymbol{C}^{m-1})$ of $\sigma$-algebras on which an $m$th-order belief $\mu_i^m$ can be defined is given by

$$\mathscr{S}_i^m(\boldsymbol{C}^{m-1}) := \Big\{\mathcal{F}_\Theta \times \{C_j^{m-1}, \emptyset\}, \mathcal{F}_\Theta \times \mathcal{F}_{j,1}^{m-1}(\boldsymbol{C}^{m-1}), \dots, \mathcal{F}_\Theta \times \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}^{m-1})\Big\},$$

where, for $\ell \leq m$, $\mathcal{F}_{j,\ell-1}^{m-1}(\boldsymbol{C}^{m-1})$ is the $\sigma$-algebra generated by the sets of the form

$$\big\{(\mu_j^1, \dots, \mu_j^{m-1}) \in C_j^{m-1} : \Sigma(\mu_j^{\ell-1}) = \mathcal{F}_\Theta \times \mathcal{F}, \mu_j^{\ell-1}(E) \geq p\big\} \tag{3.1}$$

for $\mathcal{F}_\Theta \times \mathcal{F} \in \mathscr{S}_j^{\ell-1}(\boldsymbol{C}^{\ell-2})$, $E \in \mathcal{F}_\Theta \times \mathcal{F}$ and $p \in [0, 1]$.[10] The $\sigma$-algebra $\mathcal{F}_{j,\ell-1}^{m-1}(\boldsymbol{C}^{m-1})$ contains precisely the subsets of $(m-1)$th-order belief hierarchies in $C_j^{m-1}$ that can be described in terms of the first $\ell - 1$ orders of beliefs. In other words, every event $E \subseteq C_j^{m-1}$ in this

[10]For $\ell = 2$, $\mathcal{F}_{j,\ell-1}^{m-1}(\boldsymbol{C}^{m-1})$ is the $\sigma$-algebra generated by the sets $\big\{(\mu_j^1, \dots, \mu_j^{m-1}) \in C_j^{m-1} : \mu_j^{\ell-1}(E) \geq p\big\}$ for $E \in \mathcal{F}_\Theta$ and $p \in [0, 1]$.

$\sigma$-algebra can be characterized by some restriction on the $(\ell-1)$th-order belief hierarchies: every $(m-1)$th-order belief hierarchy in $E$ satisfies that restriction, and, conversely, $E$ contains every belief hierarchy in $C_j^{m-1}$ that satisfies this restriction. Thus, the events in this $\sigma$-algebra are completely determined by player $j$'s belief up to order $\ell-1$. In particular, the events in $\mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C})$ are determined by $j$'s belief up to order $m-1$. On the other hand, the trivial $\sigma$-algebra $\{C_j^{m-1}, \emptyset\}$ does not distinguish the belief hierarchies in any way. Thus, the $\sigma$-algebras in $\mathscr{S}_i^m(\boldsymbol{C})$ form a filtration:

$$\{C_j^{m-1}, \emptyset\} \subseteq \mathcal{F}_{j,1}^{m-1}(\boldsymbol{C}) \subseteq \ldots \subseteq \mathcal{F}_{j,m-2}^{m-1}(\boldsymbol{C}) \subseteq \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}).$$

With this selection of $\sigma$-algebras, the depth of reasoning of a belief hierarchy is well-defined, as we show in Section 3.3.

Before discussing the depth of reasoning of belief hierarchies, we consider a few examples. We first consider the space of belief hierarchies constructed by Mertens and Zamir (1985) and others, in which every $m$th-order belief $\mu_i^m$ is defined on the $\sigma$-algebra that describes other player's belief up to order $m-1$, for every $m$:

**Example 6. (Mertens and Zamir, 1985)** To construct this space, we make some topological assumptions; this will allow us to show that the space of belief hierarchies we construct defines a type space, which will be useful for showing that certain Harsanyi type spaces exist (Lemma 6.1). We assume that the set $\Theta$ of states of nature is Polish; examples of Polish spaces include finite and countable sets, and closed subsets of the real line (under their usual topologies). For any topological space $X$, its Borel $\sigma$-algebra is denoted by $\mathscr{B}(X)$. The set $\Delta(X, \mathscr{B}(X))$ of Borel probability measures is endowed with the topology of weak convergence; if $X$ is Polish, then so is $\Delta(X, \mathscr{B}(X))$. As is well-known, the $\sigma$-algebra $\mathcal{F}_{\Delta(X,\mathscr{B}(X))}$ coincides with the Borel $\sigma$-algebra $\mathscr{B}(\Delta(X, \mathscr{B}(X)))$ whenever $X$ is Polish.

We endow the set $\Theta$ of states of nature with its Borel $\sigma$-algebra, i.e., $\mathcal{F}_\Theta = \mathscr{B}(\Theta)$. For each $i \in N$, take $C_i^{\mathcal{U},1}$ to be the set $\Delta(\Theta, \mathcal{F}_\Theta)$ of all probability measures on $\Theta$. For $m = 2, 3, \ldots$, let $C_i^{\mathcal{U},m}$ be the set of $m$th-order belief hierarchies $(\mu_i^1, \ldots, \mu_i^m)$ (satisfying conditions $(i)$–$(iii)$ above) such that $\mu_i^m$ is defined on $\mathcal{F}_\Theta \times \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}^{\mathcal{U}})$.[11] Note that $\mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}^{\mathcal{U}}) = \mathscr{B}(C_j^{\mathcal{U},m-1})$.

By standard arguments, the spaces $C_i^{\mathcal{U},m}$ are Polish and nonempty for every $i$ and $m$. Moreover, for every $m$ and $\ell < m$, the $\sigma$-algebra $\mathcal{F}_{j,\ell-1}^{m-1}(\boldsymbol{C}^{\mathcal{U}})$ is a proper sub-$\sigma$ algebra of $\mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}^{\mathcal{U}})$. This implies that the set of $m$th-order beliefs that we include is a strict subset of the set of all $m$th-order beliefs.

[11] That is, we take $C_i^{\mathcal{U},m} := \{(\mu_i^1, \ldots, \mu_i^m) \in C_i^{\mathcal{U},m-1} \times \Delta(\Theta \times C_j^{\mathcal{U},m-1}, \mathcal{F}_\Theta \times \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}^{\mathcal{U}})) : \operatorname{marg}_{\Theta \times C_j^{\mathcal{U},m-2}} \mu_i^m = \mu_i^{m-1}\}$ for $m > 2$, and $C_i^{\mathcal{U},m} := \{(\mu_i^1, \ldots, \mu_i^m) \in C_i^{\mathcal{U},m-1} \times \Delta(\Theta \times C_j^{\mathcal{U},m-1}, \mathcal{F}_\Theta \times \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}^{\mathcal{U}})) : \operatorname{marg}_\Theta \mu_i^m = \mu_i^{m-1}\}$ for $m = 2$.

The resulting set $H_i^{\mathcal{U}} := H_i(\boldsymbol{C}^{\mathcal{U}})$ of belief hierarchies for player $i$ is the set of belief hierarchies constructed by Mertens and Zamir (1985) and others. Mertens and Zamir show that every type from a Harsanyi type space can be mapped into this space in a way that preserves beliefs. We therefore refer to the belief hierarchies $(\mu_i^1, \mu_i^2, \ldots)$ in $H_i^{\mathcal{U}}$ as *Harsanyi (belief) hierarchies.* $\triangleleft$

While the space of belief hierarchies in Example 6 contains all belief hierarchies generated by types in Harsanyi type spaces, it does not contain all belief hierarchies: it does not contain belief hierarchies $(\mu_i^1, \mu_i^2, \ldots)$ for which for some $m$, the beliefs about the other player's $m$th-order beliefs are defined on a $\sigma$-algebra that only describes the other player's belief up to some order $\ell - 1 < m - 1$.

The following example constructs a space of belief hierarchies that does not have this restriction:

**Example 7.** Again, assume that $\Theta$ is a Polish space and that $\mathcal{F}_\Theta$ is its Borel $\sigma$-algebra $\mathscr{B}(\Theta)$. We endow the union $X$ of a family of topological spaces $X_z$, $z \in Z$, with the topology whose open sets are precisely the subsets $U$ of $X$ such that $U \cap X_z$ is open in $X_z$ for every $z \in Z$. Then, for any countable collection $\mathscr{S}$ of $\sigma$-algebras on a Polish space $X$, the union $\Delta(X, \mathscr{S})$ of spaces $\Delta(X, \mathcal{F})$, $\mathcal{F} \in \mathscr{S}$, is Polish (e.g., Kechris, 1995, Prop. 3.3).

For each $i \in N$, let $C_i^{*,1} = \Delta(\Theta, \mathcal{F}_\Theta)$ be the set of all probability measures on $\Theta$, as before. For $m = 2, 3, \ldots$, let $C_i^{*,m}$ be the set of $m$th-order belief hierarchies $(\mu_i^1, \ldots, \mu_i^m)$ (satisfying conditions $(i)$–$(iii)$ above) such that $\mu_i^m$ is defined on any of the $\sigma$-algebras in $\mathscr{S}_i^m(\boldsymbol{C}^*)$.[12] Again, the $\sigma$-algebra $\mathcal{F}_{j,\ell-1}^{k-1}(\boldsymbol{C}^*)$ is a proper sub-$\sigma$ algebra of $\mathcal{F}_{j,k-1}^{k-1}(\boldsymbol{C}^*)$ for every $\ell < k$. By standard arguments, the space $C_i^{*,m}$ of $m$th-order belief hierarchies is nonempty and Polish.

The resulting set $H_i^* := H_i(\boldsymbol{C}^*)$ of belief hierarchies contains the space $H_i^{\mathcal{U}}$ of all Harsanyi hierarchies by construction. The set $H_i^*$ additionally contains belief hierarchies $(\mu_i^1, \mu_i^2, \ldots)$ such that for some $k < \infty$, for every $m \leq k$, the belief about $j$'s $(m-1)$th-order beliefs is defined on the $\sigma$-algebra $\mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}^*)$ that describes $j$'s $(m-1)$th-order beliefs, while for $m > k$, it is defined on the $\sigma$-algebra $\mathcal{F}_{j,k-1}^{m-1}(\boldsymbol{C}^*)$ that describes $j$'s $(k-1)$th-order beliefs. Thus, the set $H_i^*$ strictly contains the set $H_i^{\mathcal{U}}$ of Harsanyi hierarchies. See the online appendix for further details. $\triangleleft$

Before turning to the depth of reasoning of belief hierarchies, let us note that the typical approach in the literature is to construct a space of belief hierarchies that contains all belief hierarchies in some sense, that is, to construct a so-called universal space. Instead, we define a

[12] That is, we take $C_i^{*,m} := \{(\mu_i^1, \ldots, \mu_i^m) \in C_i^{*,m-1} \times \Delta(\Theta \times C_j^{*,m-1}, \mathscr{S}_i^m(\boldsymbol{C}^*)) : \operatorname{marg}_{\Theta \times C_j^{*,m-2}} \mu_i^m = \mu_i^{m-1}\}$ for $m > 2$, and $C_i^{*,m} := \{(\mu_i^1, \ldots, \mu_i^m) \in C_i^{*,m-1} \times \Delta(\Theta \times C_j^{*,m-1}, \mathscr{S}_i^m(\boldsymbol{C}^*)) : \operatorname{marg}_\Theta \mu_i^m = \mu_i^{m-1}\}$ for $m = 2$.

family of spaces of belief hierarchies, by varying $\boldsymbol{C}$. For the Harsanyi case, the two approaches are equivalent (under certain topological restrictions). For the present setting, this is not the case; see Section 7.

### 3.3. Depth of reasoning

We define the depth of reasoning of a belief hierarchy $h_i = (\mu_i^1, \mu_i^2, \ldots)$ to be infinite if for every $m$, the induced $m$th-order belief $\mu_i^m$ can assign a probability to all events that are expressible in terms of player $j$'s $m$th-order beliefs. The belief hierarchy has a finite depth of reasoning $k$ if its induced $m$th-order belief can assign a probability only to events that can be expressed in terms of player $j$'s beliefs of order at most $k-1$. Formally:

**Definition 1.** Let $h_i = (\mu_i^1, \mu_i^2, \ldots) \in H_i(\boldsymbol{C})$ be a belief hierarchy. Then:

- $h_i$ has *infinite depth*, denoted $d_i^{\boldsymbol{C}}(h_i) = \infty$, if $\mu_i^m$ is a probability measure on $\mathcal{F}_\Theta \times \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C})$ for all $m = 1, 2, \ldots$;
- $h_i$ has *finite depth* $k = 1, 2, \ldots$, denoted $d_i^{\boldsymbol{C}}(h_i) = k$, if the following hold:
  - for each $m \leq k$, $\mu_i^m$ is a probability measure on $\mathcal{F}_\Theta \times \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C})$;
  - for each $m > k$, $\mu_i^m$ is a probability measure on $\mathcal{F}_\Theta \times \mathcal{F}_{j,k-1}^{m-1}(\boldsymbol{C})$, and

$$\mathcal{F}_{j,k-1}^{m-1}(\boldsymbol{C}) \subsetneq \mathcal{F}_{j,k}^{m-1}(\boldsymbol{C}) \subseteq \cdots \subseteq \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}).$$

By construction, the depth of reasoning of a belief hierarchy is well-defined:[13]

**Lemma 3.1.** For any belief hierarchy $h_i = (\mu_i^1, \mu_i^2, \ldots) \in H_i(\boldsymbol{C})$, there is a unique $k = \infty, 1, 2, \ldots$ such that $d_i^{\boldsymbol{C}}(h_i) = k$.

Intuitively, the $\sigma$-algebras in $\mathscr{S}_i^m(\boldsymbol{C})$, $m = 2, 3, \ldots$ are chosen in such a way that for each $m$th-order belief $\mu_i^m$, there is some $k \leq m$ such that $\mu_i^m$ can assign a probability to precisely those events that can be expressed in terms of the other player's $(k-1)$th-order beliefs. The coherency condition $(ii)$ then ensures that the depth of a belief hierarchy is well-defined: if $\mu_i^m$ can assign a probability only to order-$(k-1)$ events for $k < m$, then $\mu_i^{m+1}$ can assign a probability only to order-$(k-1)$ events.

[13] While the depth of reasoning $d_i^{\boldsymbol{C}}(h_i)$ of a belief hierarchy $h_i \in H_i(\boldsymbol{C})$ is defined relative to $\boldsymbol{C}$, there is a direct relationship between the depth of reasoning of different belief hierarchies: for any $\boldsymbol{C}$, there is a measurable embedding of $H_i(\boldsymbol{C})$ into the set $H_i^*$ of belief hierarchies constructed in Example 7, and any two belief hierarchies $h_i \in H_i(\boldsymbol{C})$ and $h_i' \in H_i(\boldsymbol{C}')$ (potentially from different spaces) that are mapped into the same belief hierarchy $h_i^* \in H_i^*$ have the same depth of reasoning as $h_i^*$.

Going back to the examples in the previous section, we see that the belief hierarchies in $H_i^{\mathcal{U}}$ in Example 6 all have an infinite depth of reasoning. By contrast, the belief hierarchies in $H_i^*$ in Example 7 can have any depth of reasoning: for every $k = 1, 2, \dots, \infty$, there are hierarchies that have depth $k$. In addition, $H_i^*$ also contains belief hierarchies of infinite depth that assign positive probability to belief hierarchies of finite depth, and so on; see the online appendix for details.

# 4. Type spaces

While the construction of players' belief hierarchies in the previous section allows us to directly model their higher-order beliefs, including their depth of reasoning, it would be desirable to have a model that does not require us to write out the belief hierarchies explicitly, in the vein of the type spaces introduced by Harsanyi (1967–1968). In this section, we generalize the concept of a Harsanyi type space. As we show in Section 5, each type induces a belief hierarchy, just like Harsanyi types do, except that the belief hierarchy can be of finite depth.

A ($\Theta$-based) *type space* is a tuple

$$(T_i, \mathcal{S}_i, \Sigma_i, \beta_i)_{i \in N}$$

that satisfies Assumption 1 below. For each player $i$, $T_i$ is a nonempty set of *types*, and $\mathcal{S}_i$ is a nonempty collection of $\sigma$-algebras on $T_i$. The function $\Sigma_i$ maps the types in $T_i$ to a $\sigma$-algebra $\Sigma_i(t_i) \in \mathcal{S}_j$ on $T_j$, and $\beta_i$ maps each type $t_i$ into a *belief* $\beta_i(t_i) \in \Delta(\Theta \times T_j, \mathcal{F}_\Theta \times \Sigma_i(t_i))$. We refer to $\beta_i$ as player $i$'s *belief map*.

Assumption 1 imposes some further restrictions on the $\sigma$-algebras in $\mathcal{S}_i$, $i \in N$, to ensure that each type generates a well-defined belief hierarchy. Thus, this assumption plays a similar role as the familiar condition in the definition of Harsanyi type spaces that belief maps be measurable, as we discuss below. To state the assumption, we need some more definitions. We say that a $\sigma$-algebra $\mathcal{F}_i$ on the type set $T_i$ of player $i$ *dominates* a $\sigma$-algebra $\mathcal{F}_j$ on the type set $T_j$ of player $j$ if for every event $E \in \mathcal{F}_\Theta \times \mathcal{F}_j$ and $p \in [0, 1]$,

$$\left\{t_i \in T_i : E \in \mathcal{F}_\Theta \times \Sigma_i(t_i), \beta_i(t_i)(E) \geq p\right\} \in \mathcal{F}_i.$$

If $\mathcal{F}_i$ dominates $\mathcal{F}_j$, then we write $\mathcal{F}_i \succ \mathcal{F}_j$; if $\mathcal{F}_i$ is the coarsest $\sigma$-algebra that dominates $\mathcal{F}_j$, we write $\mathcal{F}_i \succ^* \mathcal{F}_j$. (The coarsest $\sigma$-algebra that dominates $\mathcal{F}_j$ exists: it is the $\sigma$-algebra that is the intersection of all $\sigma$-algebras on $T_i$ that dominate $\mathcal{F}_j$.) Two $\sigma$-algebras $\mathcal{F}_i$ and $\mathcal{F}_j$ on $T_i$ and $T_j$, respectively, that dominate each other will be called a *mutual-dominance pair*. We are now ready to state the condition:

**Assumption 1.** For every player $i \in N$ and any $\sigma$-algebra $\mathcal{F}_i \in \mathcal{S}_i$ such that $\mathcal{F}_i \neq \{T_i, \emptyset\}$, there is a $\sigma$-algebra $\mathcal{F}_j \in \mathcal{S}_j$ such that one of the following holds:

(a) $(\mathcal{F}_i, \mathcal{F}_j)$ is a mutual-dominance pair; or

(b) $\mathcal{F}_i$ is the coarsest $\sigma$-algebra that dominates $\mathcal{F}_j$, i.e., $\mathcal{F}_i \succ^* \mathcal{F}_j$.

It follows immediately that each Harsanyi type space is a type space in our sense. Recall that a ($\Theta$-based) *Harsanyi type space* is a tuple $\mathcal{T}^{\mathcal{H}} = (T_i^{\mathcal{H}}, \beta_i^{\mathcal{H}})_{i \in N}$, where for each player $i$, the type set $T_i^{\mathcal{H}}$ is endowed with some fixed $\sigma$-algebra $\mathcal{F}_i^{\mathcal{H}}$, and the belief maps $\beta_i^{\mathcal{H}}$ are measurable.[14] This *measurability condition* is equivalent to the assumption that the $\sigma$-algebras on the type sets form a mutual-dominance pair.[15] Hence, any Harsanyi type space $\mathcal{T}^{\mathcal{H}} = (T_i^{\mathcal{H}}, \beta_i^{\mathcal{H}})_{i \in N}$ can be viewed as a type space in our sense, and we sometimes write $\mathcal{T}^{\mathcal{H}} = (T_i^{\mathcal{H}}, \mathcal{S}_i^{\mathcal{H}}, \Sigma_i^{\mathcal{H}}, \beta_i^{\mathcal{H}})_{i \in N}$, where $\mathcal{S}_i^{\mathcal{H}} := \{\mathcal{F}_i^{\mathcal{H}}\}$ and $\Sigma_i^{\mathcal{H}}$ is the trivial mapping.

Thus, Assumption 1 relaxes the standard measurability condition for Harsanyi type spaces. Assumption 1 is in fact strictly weaker than the measurability condition: the type space in Example 2, for example, satisfies Assumption 1, but the belief maps are not measurable (with respect to the players' $\sigma$-algebras). Note that Assumption 1 is easy to verify: as with the measurability condition for Harsanyi type spaces, we only need to consider the relation between two $\sigma$-algebras. See Section 7 for further discussion.

# 5. From types to belief hierarchies

In this section, we first show that every type can be mapped into a belief hierarchy. We then discuss the depth of reasoning of types.

## 5.1. Mapping types into belief hierarchies

To map each type in a type space into a belief hierarchy, we simultaneously construct the space of belief hierarchies generated by the type space, *and* the functions that maps each type into a belief hierarchy. Essentially, we use the same construction as in Section 3.2, where we built up belief hierarchies using arbitrary subsets $C_i^m$ of $m$th-order belief hierarchies for each

[14] This specifications covers most of the alternative definitions in the literature, such as those that require that type sets be separable metrizable or Polish, and assume that the belief maps are Borel measurable or continuous.

[15] This can be seen by noting that a function $f : X \to Y$ is measurable (with respect to the $\sigma$-algebras $\mathcal{F}_X$ and $\mathcal{F}_Y$ on $X$ and $Y$, respectively) if and only if the inverse images $f^{-1}(B)$ of subsets $B \subseteq Y$ that generate $\mathcal{F}_Y$ belong to $\mathcal{F}_X$ (e.g., Aliprantis and Border, 2005, Corollary 4.24).

player $i \in N$, except that here the subsets of $m$th-order belief hierarchies are derived from the type space.

Fix a type space $\mathcal{T} = (T_i, \mathcal{S}_i, \Sigma_i, \beta_i)_{i \in N}$ and a player $i \in N$. For each player $i \in N$, we define a mapping $h_i^{\mathcal{T},1}$ from $T_i$ to $\Delta(\Theta, \mathcal{F}_\Theta)$ by $h_i^{\mathcal{T},1}(t_i) := \operatorname{marg}_\Theta \beta_i(t_i)$. Clearly, $h_i^{\mathcal{T},1}(t_i) \in \Delta(\Theta, \mathcal{F}_\Theta)$. Define $C_i^{\mathcal{T},1} := h_i^{\mathcal{T},1}(T_i)$ to be the image of $h_i^{\mathcal{T},1}$, and $\boldsymbol{C}^{\mathcal{T},1} := \prod_{n \in N} C_n^{\mathcal{T},1}$. Let $\mathcal{F}_{i,1}^{\mathcal{T},1}(\boldsymbol{C}^{\mathcal{T}})$ be the relative $\sigma$-algebra on $C_i^{\mathcal{T},1}$ induced by $\mathcal{F}_{\Delta(\Theta,\mathcal{F}_\Theta)}$.

For $m > 1$, suppose that for each player $i \in N$ and for each $\ell \leq m-1$, the spaces $C_i^{\mathcal{T},\ell}$ and $\boldsymbol{C}^{\mathcal{T},\ell} = \prod_{n \in N} C_n^{\ell}$ have been defined, and that $\mathscr{S}_i^{\ell}(\boldsymbol{C}^{\mathcal{T}})$ is a collection of $\sigma$-algebras on $\Theta \times C_j^{\mathcal{T},\ell-1}$.[16] Also, assume that the functions $h_i^{\mathcal{T},\ell}$ from $T_i$ into $C_i^{\mathcal{T},\ell}$ have been defined.[17] Define

$$\mathscr{S}_i^{m}(\boldsymbol{C}^{\mathcal{T}}) := \Big\{ \mathcal{F}_\Theta \times \{C_j^{\mathcal{T},m-1}, \emptyset\}, \mathcal{F}_\Theta \times \mathcal{F}_{j,1}^{m-1}(\boldsymbol{C}^{\mathcal{T}}), \ldots, \mathcal{F}_\Theta \times \mathcal{F}_{j,m-1}^{m-1}(\boldsymbol{C}^{\mathcal{T}}) \Big\},$$

where, for $\ell \leq m$, $\mathcal{F}_{j,\ell-1}^{m-1}(\boldsymbol{C}^{\mathcal{T}})$ is generated by the sets

$$\big\{ (\mu_j^1, \ldots, \mu_j^{m-1}) \in C_j^{\mathcal{T},m-1} : \Sigma(\mu_j^{\ell-1}) = \mathcal{F}_\Theta \times \mathcal{F}, \mu_j^{\ell-1}(E) \geq p \big\},$$

for $\mathcal{F}_\Theta \times \mathcal{F} \in \mathscr{S}_j^{\ell-1}(\boldsymbol{C}^{\mathcal{T}})$, $E \in \mathcal{F}_\Theta \times \mathcal{F}$, and $p \in [0,1]$. Then, define the mapping $h_i^{\mathcal{T},m}$ from $T_i$ to $C_i^{\mathcal{T},m-1} \times \Delta(\Theta \times C_j^{\mathcal{T},m-1}, \mathscr{S}_i^{\mathcal{T},m}(\boldsymbol{C}^{\mathcal{T}})))$ by:

$$h_i^{\mathcal{T},m}(t_i) := \big( h_i^{\mathcal{T},m-1}(t_i), \mu_i^k(t_i) \big),$$

where $\mu_i^k(t_i)$ is the $k$th-order belief induced by $t_i$, defined by

$$\mu_i^k(t_i)(E) = \beta_i(t_i) \left( \big\{ (\theta, t_j) : (\theta, h_j^{\mathcal{T},m-1}(t_j)) \in E \big\} \right)$$

for any $E \subseteq \Theta \times C_j^{\mathcal{T},m-1}$ such that this probability is well-defined. Let $C_i^{\mathcal{T},m}$ be the image of $h_i^{\mathcal{T},m}$, and write $\boldsymbol{C}^{\mathcal{T},m} := \prod_{n \in N} C_n^{\mathcal{T},m}$.

**Lemma 5.1.** For every $i \in N$ and $m = 1, 2, 3, \ldots$, the functions $h_i^{\mathcal{T},m}$ are well-defined.

The key to proving Lemma 5.1 is to relate the $\sigma$-algebra $\Sigma_i(t_i) \in \mathcal{S}_j$ of type $t_i$ on $T_j$ to the $\sigma$-algebras on the space $C_j^{\mathcal{T},m-1}$ of $(m-1)$th-order hierarchies induced by types in $T_j$. This is nontrivial, because Assumption 1 allows many different relations between the various $\sigma$-algebras on players' types; for example, we could have cycles of $\sigma$-algebras that dominate each other, or infinite chains. In the proof of Lemma 5.1, we "categorize" the $\sigma$-algebras while we are performing an induction. This gives the structure necessary to make the connection between the $\sigma$-algebras on the type space and the $\sigma$-algebras on belief hierarchies.

[16]For $\ell = 1$, we take $\mathscr{S}_i^{\ell}(\boldsymbol{C}^{\mathcal{T}})$ to be the singleton $\{\mathcal{F}_\Theta\}$.

[17]This is with some abuse of notation: the range of $h_i^{\mathcal{T},\ell}$ is in fact a superset of $C_i^{\mathcal{T},\ell}$, as can be seen below; also see Lemma 5.1.

We can now construct the space of belief hierarchies that are generated by some type in $\mathcal{T}$. The sequence $(\boldsymbol{C}^{\mathcal{T},1}, \boldsymbol{C}^{\mathcal{T},2}, \ldots)$ defines a space of belief hierarchies, i.e., it satisfies conditions $(i)$–$(iii)$ in Section 3.2. If we write $\mu_i^1(t_i)$ for the first-order belief $\operatorname{marg}_\Theta \beta_i(t_i)$ induced by $t_i$, then, for every type $t_i \in T_i$,

$$h_i^{\mathcal{T}}(t_i) = \big(\mu_i^1(t_i), \mu_i^2(t_i), \ldots\big)$$

is a belief hierarchy. We refer to $h_i^{\mathcal{T}}(t_i)$ as the *belief hierarchy induced (or generated) by* $t_i$.

We thus have the following result:

**Theorem 5.2.** For every type space $\mathcal{T} = (T_i, \mathcal{S}_i, \Sigma_i, \beta_i)_{i \in N}$, and for each player $i \in N$, the belief hierarchies in $H_i(\boldsymbol{C}^{\mathcal{T}})$ are precisely those that are generated by the types in $T_i$. That is,

- for each type $t_i \in T_i$, there is a belief hierarchy $(\mu_i^1, \mu_i^2, \ldots) \in H_i(\boldsymbol{C}^{\mathcal{T}})$ such that $h_i^{\mathcal{T}}(t_i) = (\mu_i^1, \mu_i^2, \ldots)$;
- for every belief hierarchy $(\mu_i^1, \mu_i^2, \ldots) \in H_i(\boldsymbol{C}^{\mathcal{T}})$, there is a type $t_i \in T_i$ such that $(\mu_i^1, \mu_i^2, \ldots) = h_i^{\mathcal{T}}(t_i)$.

The proof follows directly from Lemma 5.1 and the fact that the construction above gives a space of belief hierarchies.

## 5.2. Depth of reasoning

By Lemma 3.1 and Theorem 5.2, each type $t_i$ generates a belief hierarchy $h_i^{\mathcal{T}}(t_i)$ of well-defined depth. With some abuse of terminology, we refer to the depth $d_i^{\boldsymbol{C}^{\mathcal{T}}}(h_i^{\mathcal{T}}(t_i))$ of reasoning of the hierarchy induced by a type $t_i$ as the depth of reasoning of the type, and write $d_i^{\mathcal{T}}(t_i)$ for $d_i^{\boldsymbol{C}^{\mathcal{T}}}(h_i^{\mathcal{T}}(t_i))$. As we discuss now, the depth of reasoning of a type can be determined directly from the type space.

The first step is to recognize that there is a tight connection between the $\sigma$-algebras on the type sets and the $\sigma$-algebras on the belief hierarchies. In the course of proving Lemma 5.1, we establish the following result, which we note for reference:

**Corollary 5.3.** For every player $i \in N$ and type $t_i \in T_i$, if $t_i$ has depth $k < \infty$, then

$$\begin{aligned} \Sigma_i(t_i) &= \left\{ \left\{ t_j \in T_j : h_j^{\mathcal{T},k-1}(t_j) \in B_j^{k-1} \right\} : B_j^{k-1} \in \mathcal{F}_{j,k-1}^{k-1}(\boldsymbol{C}^{\mathcal{T}}) \right\} \\ &\subsetneq \left\{ \left\{ t_j \in T_j : h_j^{\mathcal{T},k}(t_j) \in B_j^{k} \right\} : B_j^{k} \in \mathcal{F}_{j,k}^{k}(\boldsymbol{C}^{\mathcal{T}}) \right\}; \end{aligned} \tag{5.1}$$

otherwise, if $t_i$ has an infinite depth of reasoning, then

$$\Sigma_i(t_i) \supseteq \left\{ \left\{ t_j \in T_j : h_j^{\mathcal{T},m}(t_j) \in B_j^{m} \right\} : B_j^{m} \in \mathcal{F}_{j,m}^{m}(\boldsymbol{C}^{\mathcal{T}}) \right\} \tag{5.2}$$

for all $m$.

Thus, if type $t_i$ has depth $k < \infty$, then its $\sigma$-algebra is generated by the function that maps its opponent's types into their $(k-1)$th-order belief hierarchies. If $t_i$ has an infinite depth of reasoning, then its $\sigma$-algebra contains all subsets of types that can be distinguished on the basis of their finite-order belief hierarchies. Using Corollary 5.3, it is straightforward to show that the $\sigma$-algebra of a type $t_i$ of depth $k$ separates the types in $T_j$ if and only if these types for $j$ differ in their (induced) $(k-1)$th-order beliefs; similarly, if $t_i$ has infinite depth, then its $\sigma$-algebra separates the types in $T_j$ if the induced beliefs of these types differ at some order.[18] For future reference, we denote the $\sigma$-algebra in (5.1) by $\sigma(h_j^{\mathcal{T},k-1})$.

While the expressions in (5.1) and (5.2) refer to the hierarchy mappings $h_j^{\mathcal{T},k}$, and thus to the spaces of belief hierarchies, it is in fact possible to determine the depth of reasoning of a type directly from the $\sigma$-algebras on the type sets, without references to hierarchy mappings or belief hierarchies. For example, types from Harsanyi type spaces have an infinite depth, as should be expected:

**Observation 1. (Harsanyi type spaces)** If $\mathcal{T}^{\mathcal{H}}$ is a Harsanyi type space, and $t_i$ is a type in $\mathcal{T}^{\mathcal{H}}$, then $d_i^{\mathcal{T}^{\mathcal{H}}}(t_i) = \infty$.

The proof follows directly from Lemma 5.1, and is thus omitted. For example, as is well-known, the space $\boldsymbol{C}^{\mathcal{U}}$ of belief hierarchies constructed in Example 6 defines a type space, the so-called universal (Harsanyi) type space (e.g., Mertens and Zamir, 1985), and every type in this type space has an infinite depth of reasoning.

As a second example, it is easy to characterize the type spaces in which all types have the same finite depth $k$:

**Observation 2. (Uniform finite depth)** Fix a type space $\mathcal{T} = (T_i, \mathcal{S}_i, \Sigma_i, \beta_i)_{i \in N}$.

(a) Suppose that for each player $i \in N$, every type $t_i \in T_i$ is endowed with the same $\sigma$-algebra $\mathcal{F}_j \in \mathcal{S}_j$, and that $\mathcal{F}_a$ does not dominate $\mathcal{F}_b$ or vice versa. Then there is $k = 1, 2, \ldots$ such that for each $i \in N$, the $\sigma$-algebra $\mathcal{F}_i$ dominates exactly $k-1$ $\sigma$-algebras in $\mathcal{S}_j$, and the depth of each type equals $k$.

(b) Conversely, suppose that each type has the same finite depth. Then for each player $j \in N$, there is a $\sigma$-algebra $\mathcal{F}_j \in \mathcal{S}_j$ such that every type $t_i \in T_i$ is endowed with the $\sigma$-algebra $\Sigma_i(t_i) = \mathcal{F}_j$, and $\mathcal{F}_a$ does not dominate $\mathcal{F}_b$ or vice versa.

[18] Recall that a $\sigma$-algebra $\mathcal{F}$ on a space $X$ separates two (distinct) elements $x, x'$ of $X$ if there is a subset $B \in \mathcal{F}$ such that $x \in B$ and $x' \notin B$. As is well-known, the $\sigma$-algebra $\Sigma_i(t_i)$ of a type that has infinite depth may separate two types $t_j, t'_j$ even if the types induce the same belief hierarchy (i.e., the type space is redundant) (e.g., Mertens and Zamir, 1985).

Again, the proof follows directly from Lemma 5.1. An example of a type space with uniform finite depth is the type space in Figure 2: every type for player $i = a, b$ is endowed with the $\sigma$-algebra $\mathcal{F}_j^*$, and $\mathcal{F}_a^*$ does not dominate $\mathcal{F}_b^*$ or vice versa. Moreover, the $\sigma$-algebras $\mathcal{F}_a^*$ and $\mathcal{F}_b^*$ dominate only the trivial $\sigma$-algebra (on Bob's and Ann's type set, respectively), so that every type has depth 2.

While the observations above apply to type spaces in which every type has the same (finite or infinite) depth, there are also type spaces in which types can have different depths of reasoning, so that there can be uncertainty about a player's depth. For example, the online appendix shows that the space $\boldsymbol{C}^*$ of belief hierarchies in Example 7 defines a type space such that for every $k$, finite or infinite, there exists a type for each player with depth $k$. Also for the general case, it is possible to determine a type's depth directly from the type space, so that we do not have to write out its belief hierarchy to know its depth, essentially by counting the number of $\sigma$-algebras that the type's $\sigma$-algebra dominates (details available upon request).

# 6. Main results

## 6.1. Preliminaries

We start with some preliminaries. As in much of the paper, we focus here on the case of two players. The results extend immediately to the general case. In the remainder of the paper, we assume that the set $\Theta$ of states of nature is finite, to avoid technicalities, and we endow $\Theta$ with its usual (discrete) $\sigma$-algebra $\mathcal{F}_\Theta$.

A ($\Theta$-based) *game* is a tuple $\mathcal{G} = (S_i, u_i)_{i \in N}$, where for each player $i$, $S_i$ is a (nonempty) finite set of actions, endowed with its standard $\sigma$-algebra $\mathcal{F}_{S_i}$, and $u_i : S \times \Theta \to \mathbb{R}$ is a payoff function (where $S := \prod_i S_i$). A ($\Theta$-based) *(Bayesian) model* is a pair $(\mathcal{G}, \mathcal{T})$, where $\mathcal{G}$ is a game, and $\mathcal{T}$ is a type space.

For simplicity, we write $\Delta(S_i)$ and $\Delta(\Theta)$ for $\Delta(S_i, \mathcal{F}_{S_i})$ and $\Delta(\Theta, \mathcal{F}_\Theta)$, respectively. Also, if $\mu$ is a probability measure on a product space $X \times Y$, and $E$ is a measurable subset of $X$, we sometimes write $\mu(E)$ for $\text{marg}_X \mu(E)$.

Fix a game $\mathcal{G} = (S_i, u_i)_{i \in N}$ and a type space $\mathcal{T} = (T_i, \mathcal{S}_i, \Sigma_i, \beta_i)_{i \in N}$. A *strategy* for player $i \in N$ is a mapping $\sigma_i : T_i \to \Delta(S_i)$, with $\sigma_i(t_i)(s_i)$ the probability that type $t_i$ plays action $s_i$. The (interim) expected utility of type $t_i$ of action $s_i \in S_i$ given strategy $\sigma_j$ of the other player is given by

$$U_i(s_i, \sigma_j; t_i) := \int_{S_j \times \Theta \times T_j} u_i(s_i, s_j, \theta) \sigma_j(t_j)(s_j) d\beta_i(t_i).$$

If the expected utility $U_i(s_i, \sigma_j; t_i)$ is well-defined for each action $s_i$, we say that the strategy

$\sigma_j$ is *comprehensible for* $t_i$. A sufficient condition for $\sigma_j$ to be comprehensible for $t_i$ is that $\sigma_j$ is measurable with respect to $\Sigma_i(t_i)$ and the usual $\sigma$-algebra on $\Delta(S_j)$.[19]

We are now ready to define the equilibrium concept.

**Definition 2.** Let $\mathcal{G}$ be a game, and let $\mathcal{T}$ be a type space. A strategy $\sigma = (\sigma_i)_{i \in N}$ is a *(Bayesian-Nash) equilibrium* of the model $(\mathcal{G}, \mathcal{T})$ if for each player $i \in N$ and type $t_i \in T_i$, the following hold:

- the strategy $\sigma_j$ is comprehensible for $t_i$; and
- for each action $s_i \in S_i$ such that $\sigma_i(t_i)(s_i) > 0$,

$$U_i(s_i, \sigma_j; t_i) \geq U_i(s_i', \sigma_j; t_i)$$

  for every action $s_i' \in S_i$.

It can be checked that this is the standard Bayesian-Nash equilibrium concept when $\mathcal{T}$ is a Harsanyi type space. The second condition in Definition 2 is just the familiar best-reply condition; the first condition is needed to ensure that each type can calculate its expected payoffs if the other player follows the equilibrium strategy. This condition is standard, but, as we observed in Example 5, it can have more "bite" for finite-depth types than for Harsanyi types.

From now on, we focus primarily on type spaces in which every type has the same depth; the results extend directly to the general case, see Remark 2 below. We refer to a type space in which every player has depth $k < \infty$ as a *depth-$k$ (type) space*; also, recall that a type space in which every type has infinite depth is a Harsanyi type space. To abstract from the issue that the set of equilibria can depend on the presence of redundant types –an issue that is orthogonal to the focus of the present paper –, we restrict attention throughout this section to type spaces that are *nonredundant* in the sense that no two types generate the same belief hierarchy (cf. Mertens and Zamir, 1985). Thus, a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ is nonredundant if for each player $i$, the hierarchy mapping $h_i^{\mathcal{T}^{\mathcal{H}}}$ is injective, and a depth-$k$ space $\mathcal{T}^k$ is ($k$th-order) nonredundant if the $k$th-order hierarchy mapping $h_i^{\mathcal{T}^k,k}$ is injective. The restriction to nonredundant type spaces is not essential.

[19]The condition that a strategy $\sigma_j$ be measurable for $t_i$ is generally not necessary for it to be comprehensible for the type: a strategy $\sigma_j$ is comprehensible for $t_i$ even if it is not measurable for the type if (1) $\sigma_j$ is measurable only on a support of $\beta_i(t_i)$; or (2) for each mixed action $\alpha_i \in \Delta(S_i)$, the function $u_i(\alpha_i, \sigma_j(\cdot), \cdot) : T_j \times \Theta \to \mathbb{R}$ (where $u_i$ is extended to mixed actions in the usual way) is measurable with respect to $\Sigma_i(t_i) \times \mathcal{F}_\Theta$. The equilibrium strategies in a "report-your-$k$th-order-beliefs" game (e.g., Dekel et al., 2006) are not measurable for types of depth $k$, but satisfy (2), and are thus comprehensible for such types.

## 6.2. Strategic equivalence

Our aim is to understand whether the equilibrium behavior of types with a finite depth can be described by the standard equilibrium concept applied to Harsanyi type spaces. Formally:

**Definition 3.** Fix a depth-$k$ space $\mathcal{T}^k = (T_i, \mathcal{S}_i, \Sigma_i, \beta_i)_{i\in N}$, and let $\mathcal{T}^{\mathcal{H}} = (T_i^{\mathcal{H}}, \beta_i^{\mathcal{H}})_{i\in N}$ be a Harsanyi type space such that there is a surjective mapping $\psi_i$ from $T_i^{\mathcal{H}}$ to $T_i$ for each player $i \in N$. The type spaces $\mathcal{T}^{\mathcal{H}}$ and $\mathcal{T}^k$ are *strategically equivalent* if for each game $\mathcal{G}$, the following hold:

(1) for every equilibrium $\sigma^k = (\sigma_i^k)_{i\in N}$ of $(\mathcal{G},\, \mathcal{T}^k)$, there is a corresponding equilibrium of $(\mathcal{G},\, \mathcal{T}^{\mathcal{H}})$, that is, the strategy profile $\sigma$, with $\sigma_i = \sigma_i^k \circ \psi_i$ for $i \in N$, is an equilibrium of $(\mathcal{G},\, \mathcal{T}^{\mathcal{H}})$; and

(2) for every equilibrium $\sigma = (\sigma_i)_{i\in N}$ of $(\mathcal{G},\, \mathcal{T}^{\mathcal{H}})$, there is a corresponding equilibrium of $(\mathcal{G},\, \mathcal{T}^k)$, that is, the strategy profile $\sigma^k = (\sigma_i^k)_{i\in N}$, with $\sigma_i = \sigma_i^k \circ \psi_i$ for $i \in N$, is an equilibrium of $(\mathcal{G},\, \mathcal{T}^k)$.

If (1) holds for every game $\mathcal{G}$, we say that $\mathcal{T}^{\mathcal{H}}$ *contains* the equilibria of $\mathcal{T}^k$; and, conversely, if (2) holds for every game $\mathcal{G}$, we say that $\mathcal{T}^k$ contains the equilibria of $\mathcal{T}^{\mathcal{H}}$. Note that if the function $\psi_i$ is not surjective, then $\sigma_i^k(t_i)$, with $\sigma_i^k$ defined by $\sigma_i = \sigma_i^k \circ \psi_i$, is not defined for some types $t_i$.

As noted in Section 2, when $\mathcal{T}^k$ and $\mathcal{T}^{\mathcal{H}}$ have the same type sets (i.e., for each player $i$, we have $T_i^{\mathcal{H}} = T_i$ and $\psi_i$ is the identity function), this definition simply requires that the set of equilibria in $(\mathcal{G},\, \mathcal{T}^{\mathcal{H}})$ and $(\mathcal{G},\, \mathcal{T}^k)$ coincide for every game $\mathcal{G}$. Allowing arbitrary type sets strengthens our negative result (Theorem 6.4 below), without substantially weakening the other results.

## 6.3. Equilibrium

We first consider the question whether for a given a depth-$k$ space $\mathcal{T}^k$, there is a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ with the property that for every game $\mathcal{G}$, for any equilibrium of the finite-depth model $(\mathcal{G},\, \mathcal{T}^k)$, there is a corresponding equilibrium in the Harsanyi model $(\mathcal{G},\, \mathcal{T}^{\mathcal{H}})$. As illustrated by Example 4, this does not hold for arbitrary Harsanyi type spaces. We characterize the subclass of Harsanyi type spaces with this property.

Let $k = 1, 2, \ldots$ and fix a depth-$k$ space $\mathcal{T}^k = (T_i, \mathcal{S}_i, \Sigma_i, \beta_i)_{i\in N}$; note that there is a unique $\sigma$-algebra $\mathcal{F}_i^k \in \mathcal{S}_i$ such that $\Sigma_j(t_j) = \mathcal{F}_i^k$ for all $t_j \in T_j$ (Observation 2). We define a family of Harsanyi type spaces $\mathcal{T}^{\mathcal{H}}$ that extend $\mathcal{T}^k$ in the sense that the beliefs of each type in $\mathcal{T}^{\mathcal{H}}$ up to order $k$ are consistent with the $k$th-order beliefs of a type in $\mathcal{T}^k$.

Formally, a Harsanyi type space $\mathcal{T}^{\mathcal{H}} = (T_i^{\mathcal{H}}, \{\mathcal{F}_i^{\mathcal{H}}\}, \Sigma_i^{\mathcal{H}}, \beta_i^{\mathcal{H}})_{i\in N}$ is a *Harsanyi extension* of the depth-$k$ space $\mathcal{T}^k$ if for each player $i \in N$, there is a surjective mapping $\varphi_i : T_i^{\mathcal{H}} \to T_i$ such that:

- $\varphi_i$ is measurable (with respect to $\mathcal{F}_i^k$ and $\mathcal{F}_i^{\mathcal{H}}$); and
- for each $t_i^{\mathcal{H}} \in T_i^{\mathcal{H}}$ and $E \in \mathcal{F}_\Theta \times \mathcal{F}_j^k$, we have

$$\beta_i(\varphi_i(t_i^{\mathcal{H}}))(E) = \beta_i^{\mathcal{H}}(t_i^{\mathcal{H}})\big(\{(\theta, t_j^{\mathcal{H}}) : (\theta, \varphi_j(t_j^{\mathcal{H}})) \in E\}\big).$$

Thus, the mappings $\varphi_i$, $i \in N$, preserve the belief structure of $\mathcal{T}^k$ in a similar way as so-called type morphisms do in the context of Harsanyi type spaces (cf. Mertens and Zamir, 1985). We therefore refer to $\varphi := (\varphi_i)_{i\in N}$ as an *(extended) type morphism* (from $\mathcal{T}^{\mathcal{H}}$ to $\mathcal{T}^k$), and, with some abuse of terminology, we sometimes refer to the pair $(\mathcal{T}^{\mathcal{H}}, \varphi)$ as a Harsanyi extension of $\mathcal{T}^k$.[20] For any depth-$k$ type $t_i \in T_i$, a type $t_i^{\mathcal{H}} \in T_i^{\mathcal{H}}$ is said to be an *extension* of $t_i$ if $\varphi_i(t_i^{\mathcal{H}}) = t_i$. Note that a type in $\mathcal{T}^k$ can have multiple extensions in $\mathcal{T}^{\mathcal{H}}$. It can be shown that the Harsanyi extensions of $\mathcal{T}^k$ are precisely the Harsanyi type spaces in which the $k$th-order belief hierarchies are given by those in $\mathcal{T}^k$, and there is common belief in that event. Hence, the current definition coincides with the definition given in Section 2.

The next result shows that a Harsanyi extension exists for broad class of type spaces.

**Lemma 6.1.** Suppose that $\mathcal{T}^k$ is $k$th-order nonredundant, and that for each $i \in N$, the type set $T_i$ is Polish and that the Borel $\sigma$-algebra $\mathscr{B}(T_i)$ is generated by the $k$th-order hierarchy mapping $h_i^{\mathcal{T}^k,k}$, i.e., $\mathscr{B}(T_i) = \sigma(h_i^{\mathcal{T}^k,k})$. Then $\mathcal{T}^k$ has a Harsanyi extension.

The proof is relegated to the online appendix. The next result states that a Harsanyi type space contains the equilibria of a depth-$k$ space if and only if it is a Harsanyi extension of the depth-$k$ space:

**Theorem 6.2.** Let $\mathcal{T}^k$ be a depth-$k$ type space and let $\mathcal{T}^{\mathcal{H}}$ be a Harsanyi type space such that there is a surjective mapping $\varphi_i$ for each player $i$ from $i$'s type set in $\mathcal{T}^{\mathcal{H}}$ to her type set in $\mathcal{T}^k$. The following are equivalent:

- For every game $\mathcal{G}$ and every equilibrium $\sigma^k$ of $(\mathcal{G}, \mathcal{T}^k)$, the strategy profile $\sigma$, with $\sigma_i = \sigma_i^k \circ \varphi_i$ for $i \in N$, is an equilibrium of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$;
- $(\mathcal{T}^{\mathcal{H}}, \varphi)$ is a Harsanyi extension of $\mathcal{T}^k$.

[20] The extended type morphisms as defined here do not fully generalize the type morphism of Mertens and Zamir (1985), as the type morphism of Mertens and Zamir need not be surjective. It is straightforward to define a such a generalization, but we do not need it here.

**Proof.** The proof that $\mathcal{T}^{\mathcal{H}}$ contains the equilibria of $\mathcal{T}^k$ whenever $\mathcal{T}^{\mathcal{H}}$ is a Harsanyi extension of $\mathcal{T}^k$ uses standard techniques and is therefore relegated to the appendix. To prove the converse, fix a depth-$k$ space $\mathcal{T}^k = (T_i, \mathcal{S}_i, \Sigma_i, \beta_i)_{i\in N}$, and recall that for each $i \in N$ and $t_i \in T_i$, we have that $\Sigma_i(t_i) = \mathcal{F}_j^k$ for some $\sigma$-algebra $\mathcal{F}_j^k$ on $T_j$. Let $\mathcal{T}^{\mathcal{H}} = (T_i^{\mathcal{H}}, \{\mathcal{F}_i^{\mathcal{H}}\}, \Sigma_i^{\mathcal{H}}, \beta_i^{\mathcal{H}})_{i\in N}$ be a Harsanyi type space such that for each player $i \in N$, there is a surjective mapping $\varphi_i$ from $T_i^{\mathcal{H}}$ to $T_i$ which is measurable with respect to $\mathcal{F}_i^{\mathcal{H}}$ and $\mathcal{F}_i^k$.

Suppose that $(\mathcal{T}^{\mathcal{H}}, \varphi)$ is not a Harsanyi extension of $\mathcal{T}^k$. We claim that there is a game $\mathcal{G}$ and an equilibrium $\sigma^k$ of $(\mathcal{G}, \mathcal{T}^k)$ such that the strategy profile $\sigma$, with $\sigma_i = \sigma_i^k \circ \varphi_i$ is not an equilibrium of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$.

Since $\mathcal{T}^{\mathcal{H}}$ is not a Harsanyi extension of $\mathcal{T}^k$, there is a player $i \in N$, a type $t_i^{\mathcal{H}} \in T_i^{\mathcal{H}}$, $\theta \in \Theta$, and $B \in \mathcal{F}_j^k$ such that

$$\beta_i^{\mathcal{H}}(t_i^{\mathcal{H}})(\{(\theta, t_j^{\mathcal{H}}) : \varphi_j(t_j^{\mathcal{H}}) \in B\}) \neq \beta_i(t_i)(\theta, B) \tag{6.1}$$

where we have defined $t_i := \varphi_i(t_i^{\mathcal{H}})$. Also, it is without loss of generality to assume that $\beta_i(t_i)(\theta, B) > 0$. We consider the case where $\beta_i^{\mathcal{H}}(t_i^{\mathcal{H}})(\theta) > 0$; we treat the complementary case in the appendix.

Define the game $\mathcal{G}_y = (S_n, u_n)_{n\in N}$ as follows. For each player $n \in N$, let $S_n := \{s_n^1, s_n^2\}$. For each state $\theta' \neq \theta$ and every action profile $s \in S$, let $u_n(s, \theta') := 0$. The payoffs in state $\theta$ are given by (with $i$ the row player):

| | $s_j^1$ | $s_j^2$ |
|---|---|---|
| $s_i^1$ | $y, 0$ | 0,0 |
| $s_i^2$ | 1,0 | 1,0 |

where

$$y := 1 + \frac{\beta_i(t_i)(\theta, T_j \setminus B)}{\beta_i(t_i)(\theta, B)}.$$

Consider the strategy $\sigma_j^k$ for player $j$, defined by:

$$\sigma_j^k(t_j) := \begin{cases} s_j^1 & \text{if } t_j \in B; \\ s_j^2 & \text{otherwise.} \end{cases}$$

Then, the strategy profile $\sigma_j^k$ is comprehensible for every type $t_i' \in T_i$, and type $t_i$ is indifferent between $s_i^1$ and $s_i^2$. Define the strategies $\sigma_i^k$ and $\tilde{\sigma}_i^k$ as follows. Let $\sigma_i^k(t_i)$ and $\tilde{\sigma}_i^k(t_i)$ assign probability 1 to $s_i^1$ and $s_i^2$, respectively, and for $t_i' \neq t_i$, let $\sigma_i^k(t_i')$ and $\tilde{\sigma}_i^k(t_i')$ assign probability 1 to $s_i^1$ if

$$\beta_i(t_i')(\theta, B)(y-1) - \beta_i(t_i')(\theta, T_j \setminus B) \geq 0,$$

and probability 1 to $s_i^2$ otherwise. Then all types choose a best response under $\sigma_j^k$, $\sigma_i^k$, and $\tilde{\sigma}_i^k$, and each of these strategies is comprehensible for the types of the other player. Hence, the strategy profiles $(\sigma_i^k, \sigma_j^k)$ and $(\tilde{\sigma}_i^k, \sigma_j^k)$ are equilibria of $(\mathcal{G}, \mathcal{T}^k)$.

Now consider the case where beliefs are given by $\mathcal{T}^{\mathcal{H}}$. If $\{(\theta, t_j^{\mathcal{H}}) : \varphi_j(t_j^{\mathcal{H}}) \in B\} \notin \mathcal{F}_j^{\mathcal{H}}$, then $\sigma_j = \sigma_j^k \circ \varphi_j$ is not comprehensible for $t_i^{\mathcal{H}}$, so the strategy profiles in $\mathcal{T}^{\mathcal{H}}$ corresponding to $(\sigma_i^k, \sigma_j^k)$ and $(\tilde{\sigma}_i^k, \sigma_j^k)$ are not an equilibrium of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$.

So suppose $\{(\theta, t_j^{\mathcal{H}}) : \varphi_j(t_j^{\mathcal{H}}) \in B\} \in \mathcal{F}_j^{\mathcal{H}}$. If $\beta_i^{\mathcal{H}}(t_i^{\mathcal{H}})(\{(\theta, t_j^{\mathcal{H}}) : \varphi_j(t_j^{\mathcal{H}}) \in B\}) > \beta_i(t_i)(\theta, B)$, then $(\tilde{\sigma}_i^k, \sigma_j^k)$ is not an equilibrium of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$; otherwise, the strategy profile $(\sigma_i^k, \sigma_j^k)$ is not an equilibrium of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$. $\square$

We next ask whether for a given depth-$k$ space $\mathcal{T}^k$, there is a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ such that $\mathcal{T}^k$ contains the equilibria of $\mathcal{T}^{\mathcal{H}}$. Given that we are interested in finding Harsanyi type spaces that are strategically equivalent to $\mathcal{T}^k$, Theorem 6.2 allows us to restrict attention to Harsanyi extensions of $\mathcal{T}^k$. (However, our results do not depend on this restriction.)

The next result shows that it is without loss of generality to restrict attention to order-$k$ extensions for our purposes, where an order-$k$ extension is a Harsanyi extension in which the higher-order beliefs of each type are completely determined by its $k$th-order beliefs. Formally, a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$ is of *order* $k$ if the following conditions hold:

- for each player $i \in N$ and pair of types $t_i^{\mathcal{H}}, \tilde{t}_i^{\mathcal{H}}$ in $\mathcal{T}^{\mathcal{H}}$, we have that $h_i^{\mathcal{T}^{\mathcal{H}},k}(\tilde{t}_i^{\mathcal{H}}) = h_i^{\mathcal{T}^{\mathcal{H}},k}(t_i^{\mathcal{H}})$ implies $h_i^{\mathcal{T}^{\mathcal{H}}}(\tilde{t}_i^{\mathcal{H}}) = h_i^{\mathcal{T}^{\mathcal{H}}}(t_i^{\mathcal{H}})$; and
- there is a player $i \in N$ and a pair of types $t_i^{\mathcal{H}}, \tilde{t}_i^{\mathcal{H}}$ in $\mathcal{T}^{\mathcal{H}}$ such that $h_i^{\mathcal{T}^{\mathcal{H}},k-1}(\tilde{t}_i^{\mathcal{H}}) = h_i^{\mathcal{T}^{\mathcal{H}},k-1}(t_i^{\mathcal{H}})$ and $h_i^{\mathcal{T}^{\mathcal{H}}}(\tilde{t}_i^{\mathcal{H}}) \neq h_i^{\mathcal{T}^{\mathcal{H}}}(t_i^{\mathcal{H}})$.

The first condition says that beliefs are determined completely by players' $k$th-order beliefs; the second says that beliefs are not determined by the $(k-1)$th-order beliefs of players. This ensures that the order of a Harsanyi type space is well-defined. An *order-$k$ (Harsanyi) extension* of a depth-$k$ space $\mathcal{T}^k$ is a Harsanyi extension of $\mathcal{T}^k$ that is of order $k$.[21]

**Lemma 6.3.** Let $\mathcal{T}^k$ be a depth-$k$ space that is $k$th-order nonredundant, and let $(\mathcal{T}^{\mathcal{H}}, \varphi)$ be a Harsanyi extension of $\mathcal{T}^k$. If for every game $\mathcal{G}$, for every equilibrium $\sigma$ of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$, the strategy profile $\sigma^k$, with $\sigma_i = \sigma_i^k \circ \varphi_i$ for $i \in N$, is an equilibrium of $(\mathcal{G}, \mathcal{T}^k)$, then $\mathcal{T}^{\mathcal{H}}$ is an order-$k$ extension of $\mathcal{T}^k$. Moreover, if $\mathcal{T}^{\mathcal{H}}$ is nonredundant, then it is without loss of generality to take the type sets in $\mathcal{T}^{\mathcal{H}}$ to be equal to those in $\mathcal{T}^k$, i.e., $T_i^{\mathcal{H}} = T_i$ for $i \in N$.

[21] For depth-$k$ spaces with countable type sets, an order-$k$ extension can easily be shown to exist, but existence cannot be shown in general. (While every belief of a depth-$k$ type can be extended to a belief on an appropriately finer $\sigma$-algebra (under certain topological conditions), as in the proof of Lemma 6.1, the resulting belief map need not be measurable.)

We are now ready to prove our negative result. Theorem 6.4 shows that a converse of Theorem 6.2 does not hold, at least for depth-$k$ type spaces that are nontrivial in the sense that $k$th-order beliefs are potentially strategically relevant. Formally, a depth-$k$ space $\mathcal{T}^k$ is *nontrivial* if there is a player $i$, a type $t_i$ for $i$, and an event $E \subseteq T_j$ contained in the support of $\beta_i(t_i)$ with the property that there exist types $t_j, t_j' \in E$ such that $h_j^{\mathcal{T}^k,k-1}(t_j) = h_j^{\mathcal{T}^k,k-1}(t_j')$ and $h_j^{\mathcal{T}^k,k}(t_j) \neq h_j^{\mathcal{T}^k,k}(t_j')$. If a depth-$k$ space does not satisfy this condition, then each type assigns probability one to events that are not refined further by describing the other player's $k$th-order beliefs, so that only lower-order beliefs can be strategically relevant.

**Theorem 6.4.** Let $\mathcal{T}^k$ be a nontrivial depth-$k$ type space and let $(\mathcal{T}^{\mathcal{H}}, \varphi)$ be a Harsanyi extension of $\mathcal{T}^k$. Then there is a game $\mathcal{G}$ and an equilibrium $\sigma$ of the Harsanyi model $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$ such that $\sigma^k$, with $\sigma_i = \sigma_i^k \circ \varphi_i$ for $i \in N$, is not an equilibrium of $(\mathcal{G}, \mathcal{T}^k)$.

**Proof.** Since we restrict attention to nonredundant type spaces, by Lemma 6.3, it is without loss of generality to consider order-$k$ extensions $\mathcal{T}^{\mathcal{H}}$ of $\mathcal{T}^k$ in which every player has the same type set as in $\mathcal{T}^k$. We show that there is a game $\mathcal{G}$ and a strategy profile $\sigma$ such that for every such Harsanyi extension $\mathcal{T}^{\mathcal{H}}$, the strategy profile $\sigma$ is an equilibrium of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$, but it is not an equilibrium of $(\mathcal{G}, \mathcal{T}^k)$.

We focus on the case $k \geq 2$; the proof for the case $k = 1$ is similar and can be found in the online appendix. Let $i \in N$. By Corollary 5.3, the $\sigma$-algebra of each type $t_i \in T_i$ is generated by the $(k-1)$th-order hierarchy mapping $h_j^{\mathcal{T}^k,k-1}$, i.e., $\Sigma_i(t_i) = \sigma(h_j^{\mathcal{T}^k,k-1})$ (Eq. (5.1)). Moreover, the $\sigma$-algebra $\sigma(h_j^{\mathcal{T}^k,k-1})$ is a proper sub-$\sigma$ algebra of $\sigma(h_j^{\mathcal{T}^k,k})$. Hence, there are types $t_i, t_i' \in T_i$ that differ only in their $k$th-order beliefs, that is, $h_i^{\mathcal{T}^k,k-1}(t_i) = h_i^{\mathcal{T}^k,k-1}(t_i')$ and $h_i^{\mathcal{T}^k,k}(t_i) \neq h_i^{\mathcal{T}^k,k}(t_i')$. As $\mathcal{T}^k$ is nontrivial, it is without loss of generality to assume that $t_i$ and $t_i'$ belong to the support of the belief of some type for $j$.

It follows that there exist $\theta \in \Theta$ and $B \in \sigma(h_j^{\mathcal{T}^k,k-1}) \setminus \sigma(h_j^{\mathcal{T}^k,k-2})$ such that

$$\beta_i(t_i)(\theta, B) > 0, \qquad \beta_i(t_i')(\theta, B) \neq \beta_i(t_i)(\theta, B).$$

Without loss of generality, assume that $\beta_i(t_i)(\theta, B) - \beta_i(t_i')(\theta, B) = \varepsilon$ for some $\varepsilon > 0$. Note that for any Harsanyi extension $\mathcal{T}^{\mathcal{H}}$ (with $T_i^{\mathcal{H}} = T_i$ and $\varphi_i$ the identity function for every $i \in N$), we have that $\beta_i^{\mathcal{H}}(t_i)(\theta, B) = \beta_i(t_i)(\theta, B)$, and likewise for $t_i'$.

Consider the following game, denoted $\mathcal{G}_x$. Each player $n$ has two actions, denoted by $s_n^1$ and $s_n^2$. Payoffs are given by:

| | $s_j^1$ | $s_j^2$ |
|---|---|---|
| $s_i^1$ | $x, 0$ | $0, 0$ |
| $s_i^2$ | $1, 1$ | $1, 1$ |

$\theta$

| | $s_j^1$ | $s_j^2$ |
|---|---|---|
| $s_i^1$ | $0, 0$ | $0, 0$ |
| $s_i^2$ | $1, 1$ | $1, 1$ |

$\theta' \neq \theta$

where

$$x = \frac{1}{\beta_i(t_i)(\theta, B)} + \varepsilon.$$

Clearly, the model $(\mathcal{G}_x, \mathcal{T}^k)$ has an equilibrium in which every type $t_i$ of player $i$ plays $s_i^2$ (with probability 1), and every type $t_j$ of player $j$ plays $s_j^2$; by Theorem 6.2, any model $(\mathcal{G}_x, \mathcal{T}^{\mathcal{H}})$ such that $\mathcal{T}^{\mathcal{H}}$ is a Harsanyi extension of $\mathcal{T}^k$ has a corresponding equilibrium.

We show that there is another strategy profile $\sigma = (\sigma_i, \sigma_j)$ that is an equilibrium of $(\mathcal{G}_x, \mathcal{T}^{\mathcal{H}})$ for any Harsanyi extension $\mathcal{T}^{\mathcal{H}}$ (with $T_i^{\mathcal{H}} = T_i$ for every $i \in N$) which is not an equilibrium of the depth-$k$ model $(\mathcal{G}_x, \mathcal{T}^k)$. Define the strategy $\sigma_j$ for player $j$ by

$$\begin{aligned} \sigma_j(t_j)(s_j^1) &= 1 \qquad \text{if } t_j \in B; \\ \sigma_j(t_j)(s_j^2) &= 1 \qquad \text{if } t_j \notin B. \end{aligned}$$

Note that $\sigma_j$ is comprehensible for each type of player $i$. Then the difference in expected payoffs for $t_i$ between $s_i^1$ and $s_i^2$ is

$$U_i(s_i^1, \sigma_j; t_i) - U_i(s_i^2, \sigma_j; t_i) = \beta_i(t_i)(\theta, B)x - 1.$$

Likewise, the difference in expected payoffs for $t_i'$ between $s_i^1$ and $s_i^2$ is

$$U_i(s_i^1, \sigma_j; t_i') - U_i(s_i^2, \sigma_j; t_i') = \beta_i(t_i')(\theta, B)x - 1.$$

It can be verified that type $t_i$ strictly prefers $s_i^1$, and type $t_i'$ strictly prefers $s_i^2$. If we set $\sigma_i(t_i)(s_i^1) = 1$, $\sigma_i(t_i')(s_i^2) = 1$, and for $\tilde{t}_i \neq t_i, t_i'$, we define $\sigma_i(\tilde{t}_i)(s_i^1) = 1$ if $U_i(s_i^1, \sigma_j; \tilde{t}_i) - U_i(s_i^2, \sigma_j; \tilde{t}_i) \geq 0$, and $\sigma_i(\tilde{t}_i)(s_i^2) = 1$ otherwise, then $\sigma = (\sigma_i, \sigma_j)$ is an equilibrium of $(\mathcal{G}_x, \mathcal{T}^{\mathcal{H}})$ for any Harsanyi extension $\mathcal{T}^{\mathcal{H}}$ (with $T_i^{\mathcal{H}} = T_i$ for every $i \in N$). But $\sigma$ is not an equilibrium of $(\mathcal{G}_x, \mathcal{T}^k)$, as $t_i$ and $t_i'$ differ only in their $k$th-order beliefs and play different actions. Thus, in $\mathcal{T}^k$, the strategy $\sigma_i$ is not comprehensible for the types for $j$ (regardless of the strategies chosen by the types $\tilde{t}_i \neq t_i, t_i'$), and indeed the expected utility is not well-defined for some types for $j$ in $\mathcal{T}^k$. □

The intuition behind Theorem 6.4 is that there is a tension between the conditions in Definition 2 that players choose a best response given their type, and that strategies be comprehensible. Because there are types in a depth-$k$ space that have different $k$th-order beliefs (but have the same $(k-1)$th-order beliefs), as demonstrated in Corollary 5.3, it may be optimal for these types to take different actions, given the other player's strategy. But this violates the condition that equilibrium strategies be comprehensible.

The following result then follows directly from Theorems 6.2 and 6.4:

**Corollary 6.5.** Fix $k < \infty$ and a nontrivial depth-$k$ type space $\mathcal{T}^k$. Then there is no Harsanyi type space that is strategically equivalent to $\mathcal{T}^k$.

**Remark 1.** The proofs of Theorems 6.2 and 6.4 use games in which players are indifferent among their actions. This makes it possible to clearly bring out the intuition behind the results. One could prove the same results, however, using games in which players are not indifferent. $\triangleleft$

**Remark 2.** While we have restricted attention to depth-$k$ spaces here for simplicity, Theorems 6.2 and 6.4 (and Corollary 6.5) directly extend to arbitrary type spaces, where players can be uncertain about the depth of reasoning of their opponent (as in, e.g., Example 7). $\triangleleft$

## 6.4. Rationalizability

Results analogous to Theorems 6.2 and 6.4 can be derived for a finite-depth version of (interim correlated) rationalizability (Dekel et al., 2007). Fix a depth-$k$ space $\mathcal{T}^k$ and a game $\mathcal{G}$. For each player $i \in N$ and type $t_i \in T_i$, define $R_i^{\mathcal{G},\mathcal{T}^k,0}(t_i) := S_i$, and for $m > 0$, define

$$R_i^{\mathcal{G},\mathcal{T}^k,m}(t_i) := \left\{ s_i \in S_i : \begin{array}{ll} & \text{there is } \sigma_j : \Theta \times T_j \to \Delta(S_j) \text{ s.t.} \\ (1) & \sigma_j \text{ is measurable w.r.t. } \Sigma_i(t_i); \\ (2) & \sigma_j(\theta, t_j)(s_j) > 0 \text{ implies that } s_j \in R_j^{\mathcal{G},\mathcal{T}^k,m-1}(t_j); \\ (3) & s_i \in \arg\max_{s_i' \in S_i} \int u_i(s_i', s_j, \theta)\sigma_j(\theta, t_j)(s_j)\beta_i(t_i); \end{array} \right\}.$$

to be the set of best replies for $t_i$ to the $(m-1)$th-order rationalizable actions of player $j$. As in the Harsanyi case, define the set of (interim correlated) rationalizable actions for type $t_i$ by $R_i^{\mathcal{G},\mathcal{T}^k}(t_i) := \bigcap_m R_i^{\mathcal{G},\mathcal{T}^k,m}(t_i)$.[22] Define strategic equivalence (with respect to rationalizable behavior) in the obvious way.

It is possible to show that results analogous to Theorems 6.2 and 6.4 hold, so that we have the following analogue of Corollary 6.5:

**Corollary 6.6.** Fix $k < \infty$ and a nontrivial depth-$k$ type space $\mathcal{T}^k$. Then there is no Harsanyi type space that is strategically equivalent (with respect to rationalizable behavior) to $\mathcal{T}^k$.

To understand the intuition behind Corollary 6.6, note that the profile $(R_i^{\mathcal{G},\mathcal{T}^k}(t_i))_{i \in N, t_i \in T_i}$ is the fixed point of a best-response correspondence (cf. Dekel et al., 2007, Proposition 4).

[22] While a full epistemic treatment of rationalizability, as in Battigalli et al. (2011), is beyond the scope of the present paper, we note that this definition has the following nice interpretation: a player with a finite depth $k$ has an imperfect understanding of the higher-order beliefs of the other player, but can apply the best-reply operator as many times as she wishes (since this does not require complex reasoning, just bookkeeping). This is in line with a view of rationalizability as an "algorithmic" solution concept.

Since it is a fixed point, similar issues arise as with equilibrium: the best response to a conjecture that describes the play of depth-$k$ types need not be comprehensible for depth-$k$ types. Thus, our negative results are not due to the requirement that beliefs need to be correct in equilibrium; also see Section 7.

## 6.5. Finite-order equilibrium

Our negative result (Theorem 6.4) raises the question whether it is possible to characterize the strategy profiles that form a (Bayesian-Nash) equilibrium in a game for types of a finite depth by considering a Harsanyi extension of the finite-depth type space. We define an equilibrium refinement, called finite-order equilibrium, such that for each game, the set of equilibria of a finite-depth model corresponds precisely to the set of finite-order equilibrium of the corresponding Harsanyi model.

We use that every nondegenerate Harsanyi type space that is nonredundant is the extension of some finite-depth space:

**Lemma 6.7.** Every nonredundant Harsanyi type space whose type sets contain at least two elements is a Harsanyi extension of some finite-depth space (using a straightforward generalization of the definition of a Harsanyi extension that allows the types in the finite-depth space to have different depths of reasoning).

The proof of Lemma 6.7 is straightforward, and therefore omitted. Fix a Harsanyi type space $\mathcal{T}^{\mathcal{H}}$, and suppose it is a Harsanyi extension of the finite-depth type space $\mathcal{T}$. Without loss of generality, we can take the type sets in $\mathcal{T}$ to be the same as in $\mathcal{T}^{\mathcal{H}}$. Fix a game $\mathcal{G}$, and suppose for simplicity that $\mathcal{T}^{\mathcal{H}}$ and $\mathcal{T}$ are nonredundant, and that a strategy that is comprehensible for a type in $\mathcal{T}$ is measurable with respect to the $\sigma$-algebra of that type.

Say that a strategy profile $\sigma$ is a *finite-order equilibrium* of the Harsanyi model $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$ if it is a (Bayesian-Nash) equilibrium of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$ and for each player $i$, the strategy $\sigma_i$ is measurable with respect to $i$'s $(k_j - 1)$th-order beliefs, where $k_j$ is the minimum depth of a type for player $j$ in $\mathcal{T}$.[23] It is straightforward to show that a strategy profile $\sigma$ is a finite-order equilibrium of the Harsanyi model $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$ if and only if it is an equilibrium of the finite-depth model $(\mathcal{G}, \mathcal{T})$.

That is, the equilibrium behavior of players with a finite depth of reasoning can be described with a refinement of Bayesian-Nash equilibrium that rules out equilibria that depend on players' beliefs at high order. It can be shown that there are Harsanyi models that have

[23] The reason that the measurability condition is linked to the minimum depth of the types in $\mathcal{T}$ is that the comprehensibility condition is strongest for the types of the lowest depth.

a (Bayesian-Nash) equilibrium that is not a finite-order equilibrium, even if the depth of reasoning of types is arbitrarily high. Analogous results hold for rationalizability.

The strategies that are measurable with respect to a player's $(k-1)$th-order beliefs in $\mathcal{T}^{\mathcal{H}}$ can be found using the notion of dominance (applied to $\mathcal{T}^{\mathcal{H}}$) introduced in Section 4. This means that these strategies can be identified without having to specify players' belief hierarchies. This gives a simple method to study the equilibrium behavior of players with a finite depth of reasoning that uses only Harsanyi type spaces, and that does not require modeling players' depth of reasoning or their belief hierarchies explicitly.

# 7. Discussion

**Alternative solution concepts** We have shown that Harsanyi type spaces do not capture the behavior of types with a finite depth of reasoning under (Bayesian-Hash) equilibrium or (interim correlated) rationalizability, unless one considers the appropriate refinements (Section 6.5). Since the tension between the best-response condition and comprehensibility is at the core of these results, one may ask whether it is appropriate to consider concepts like conjectural or self-confirming equilibrium (e.g., Battigalli, 1987; Dekel et al., 2004) or analogy-based expectations equilibrium (Jehiel, 2005) in the present context.

These solution concepts presume that players may lump certain states of the world together, as we do here. Unlike the present setting, there is no tension between the best-response property and a comprehensibility or measurability condition.[24] The reason is that these concept assume that a player, say, Ann, believes that the other player, say, Bob, chooses some mixture of actions in the states she lumps together, where the mixture puts positive probability only on the actions that Bob actually chooses in these states. Thus, while a player may not have correct beliefs about the actions of the other player in any particular state, her beliefs are required to be correct "on average." Since the behavior of her opponent is assumed to be constant on the set of states she lumps together, expected payoffs are well-defined, which means that there is no binding comprehensibility condition.

While it is possible to define the analogues of such concepts here, the motivation for these concepts seems weak at best in the present setting. To wit, concepts like conjectural and self-confirming equilibrium and analogy-based expectations equilibrium are have a learning motivation: players play the same game many times, and receive partial feedback on outcomes. Players lump together states of the world that they cannot distinguish based on their coarse feedback. For example, a player may learn only his own payoffs after playing the game, and

[24]These concepts do require strategies or conjectures to be measurable (and thus comprehensible), just like Bayesian-Nash equilibrium, but this condition does not have a bite as it does here.

thus lumps together states of the world that differ only in others' payoffs.

To translate this learning motivation to the present setting, we would have to assume that a player of depth $k$ receives feedback on the strategy of her opponent as a function of his $(k-1)$th-order belief, but not on his strategy as a function of his $k$th-order belief. This does not seem to be a very realistic assumption. It is an open question whether it is possible to find a motivation for such solution concepts that relies on less problematic assumptions about player feedback.

**Assumption 1 revisited.** As noted in Section 4, Assumption 1 relaxes the condition in the definition of Harsanyi type spaces that belief maps be measurable. Assumption 1 plays an important role in the characterization of the $\sigma$-algebra of each type in terms of the hierarchy mappings in the proof of Lemma 5.1, which plays a central role in the paper. To see how, note that every $\sigma$-algebra $\mathcal{F}_i^1$ on $T_i$ that dominates the trivial $\sigma$-algebra $\{T_j, \emptyset\}$ contains the subsets of $i$'s types that can be distinguished on the basis of their beliefs about $\theta$, i.e., in terms of their first-order belief hierarchies; in turn, every $\sigma$-algebra $\mathcal{F}_j^2$ on $T_j$ that dominates $\mathcal{F}_i^1$ contains the subsets of types for $j$ that are can be distinguished on the basis of their beliefs about player $i$'s first-order belief hierarchies, i.e., in terms of their second-order belief hierarchies, and so on. Assumption 1 ensures that for each type $t_i$, either its $\sigma$-algebra $\Sigma_i(t_i)$ is part of a finite chain

$$\Sigma_i(t_i) \succ^* \mathcal{F}_i^\ell \succ^* \mathcal{F}_j^{\ell-1} \succ^* \cdots \succ^* \mathcal{F}_m^1 \succ^* \{T_n, \emptyset\}$$

of $\sigma$-algebras that are the coarsest $\sigma$-algebras that dominate each other (where $n \in N$ and $m \neq n$), so that $t_i$ has a finite depth; or $\Sigma_i(t_i)$ dominates all $\sigma$-algebras (on $T_i$) that are part of such chains, in which case $t_i$ has an infinite depth.

The requirement that the $\sigma$-algebra of each type is either part of a finite chain as the one above, or dominates all such chains is essentially equivalent to Assumption 1 (in the sense that the same spaces of belief hierarchies can be modeled under both conditions). Since Assumption 1 requires one to consider only the relation between two $\sigma$-algebras at the time, just like the measurability condition, instead of requiring one to construct chains of $\sigma$-algebras of arbitrary length, we have chosen the present formulation.

Finally, we note that it is possible to relax the requirement that the $\sigma$-algebras in the chains be the coarsest to dominate each other, and just require that they dominate each other. Under this weaker condition, $\sigma$-algebras corresponding to a finite depth may include events that are unrelated to a type's depth of reasoning (i.e., events that cannot be described in terms of the hierarchy mappings). In that case, it may not be possible to compare two types of a given player that have the same depth (in a given type space) in terms of the events that they can reason about. Further study of such phenomena and their strategic implications is beyond the scope of the present paper.

**A universal space?** For the present purpose of studying equilibrium behavior, it is not necessary to construct a universal type space for our class of type spaces, as Mertens and Zamir (1985) and others have done for Harsanyi type spaces. One might nevertheless wonder whether such a type space, which embeds all other type spaces (in the sense that there is a unique type morphism from each type space into the space), can be constructed for the class of type spaces that we consider. One observation is that for any space $\boldsymbol{C}$ of belief hierarchies, there is a measurable mapping from each type space into the type space in Example 7. However, the images of such mappings in $\prod_i H_i^*$ will generally not form a belief-closed subset, as is the case for Harsanyi type spaces (Mertens and Zamir, 1985). We leave a full exploration of such issues for future research.

# 8. Related literature

**Level-$k$ and cognitive hierarchy models.** An important literature in experimental and behavioral economics studies the behavior of players with a finite depth of reasoning, where it is assumed that a player of depth $k < \infty$ applies the best-response operator $k$ times: a level-0 player is non-strategic and follows some exogenously specified strategy, while for $k > 0$, a level-$k$ player chooses a best response to a belief that his opponents are of a lower level; see footnote 1 for references, and see Crawford et al. (2012) for an excellent survey. We depart from this literature in two ways.

First, while much of the literature in experimental and behavioral economics has focused on games with complete information about payoffs,[25] we study games with incomplete information. This is of interest because in games with incomplete information, limitations on a players' depth of reasoning can affect behavior even beyond the initial periods of play. While in games with complete information, there is no remaining higher-order uncertainty once play has converged to equilibrium (as it often does in experiments), higher-order uncertainty con-

[25]Notable exceptions include Brocas et al. (2009), Crawford and Iriberri (2007), and Rogers et al. (2009), who present behavioral models for specific classes of games with incomplete information and test these models experimentally. However, these authors do not develop a theoretical framework to analyze behavior in games beyond the specific applications that they consider, like we do in the present paper. Strzalecki (2009) introduces type spaces to model uncertainty about other players' depth of reasoning in games with complete information. While his framework can be used to analyze games with incomplete information in the agent-normal form (when type sets are countable), his type spaces only allow for uncertainty about players' depth of reasoning, not about payoffs. Heifetz and Kets (2013) define a framework based on the framework presented here to study robustness questions.

tinues to play a role in games with incomplete information, so that the behavior of players with a finite depth of reasoning may differ from that of players with an infinite depth of reasoning even if players have ample experience in playing the game.

Second, while the existing literature has focused on nonequilibrium concepts, we consider an extension of Bayesian-Nash equilibrium. This allows us to isolate the effect of the assumption that players have a finite depth of reasoning from the effect of the assumption of nonequilibrium play, and to investigate whether we can use Harsanyi type spaces and Bayesian-Nash equilibrium to describe the equilibrium behavior of players with a finite depth of reasoning.

The innovation of our approach is that we model a player's depth by the set of events she can reason about, rather than by a single number, as in the rest of the literature. This extends the notion of a small world, introduced by Savage (1954) in the context of single-person decision problems, to a strategic setting. A state (of the world) in a small world describes the possible uncertainties a decision-maker faces in less detail than a state in a larger world, by neglecting certain distinctions between states. This means that "a state of the smaller world corresponds not to one state of the larger, but to a *set* of states" (Savage, 1954, p. 9, emphasis added). In the present framework, a player may ignore the distinction between types for the other player that differ only in the beliefs they generate at high order, by lumping together these types into a single set in her $\sigma$-algebra. A player with a lower depth of reasoning makes fewer distinctions between states than a player with a higher depth of reasoning, and thus has a smaller world.

This approach allows us to extend the Bayesian-Nash concept to a setting with players with an infinite depth of reasoning, which does not seem to be straightforward in models where a player's depth is simply given by a number, as in Strzalecki (2009) and Heifetz and Kets (2013). In addition, the present approach can be used to show that players with a finite depth of reasoning can attain common knowledge under certain conditions, which solves an important puzzle in philosophy (Kets, 2013).

**Robustness of predictions.** While it is well-known that game-theoretic predictions can be sensitive to small changes in higher-order beliefs (see, e.g., Rubinstein, 1989; Weinstein and Yildiz, 2007; Ely and Peski, 2011, among many others), the existing literature does not address the question whether Harsanyi type spaces can be used to model the behavior of players with a finite depth; and indeed, our results do not follow from existing results. Specifically, a key insight from our analysis is that if a type has depth $k$, then there are types that differ only in their $k$th-order beliefs. We then show that for *each* finite order $k$, for *every* pair of types that differ in their $k$th-order beliefs, there is a game in which these types have distinct optimal actions. The existing literature, by contrast, shows that the behavior of *some* types is sensitive to the specification of their beliefs at *some* order, possibly very high.

**Measurable structures on type sets.** One insight of the present paper is that, by choosing the measurable sets on which a type's belief is defined, we can get types that can reason about only finitely many orders of beliefs. Indeed, a technical contribution of this paper is to formulate a condition on the type space (Assumption 1) that guarantees that the $\sigma$-algebra of a type with a finite depth $k$ lumps together precisely the types that induce belief hierarchies that coincide up to order $k-1$ (Lemma 5.1 and Corollary 5.3). The idea that a type's $\sigma$-algebra can determine its depth of reasoning fits in with a broader literature that studies how the measurable structure associated with types in Harsanyi type spaces can implicitly impose restrictions on reasoning, i.e., on belief hierarchies (e.g., Brandenburger and Keisler, 2006; Friedenberg and Meier, 2012); see Friedenberg and Keisler (2011) for an excellent discussion and further references.

# Appendix A Proofs for Sections 3–5

## A.1 Proof of Lemma 3.1

The result follows directly from the coherency condition (*ii*). If $\mu_i^{k-1}$ is defined on the $\sigma$-algebra $\mathcal{F}_\Theta \times \mathcal{F}_{j,\ell-2}^{k-2}(\boldsymbol{C})$ for $\ell < k$, then any probability measure $\mu_i^k$ in $\Delta(\Theta \times C_j^{k-1}, \mathscr{S}_i^k(\boldsymbol{C}))$ that satisfies (*ii*) (i.e., is such that $\operatorname{marg}_{\Theta\times C_j^{k-2}} \mu_i^k = \mu_i^{k-1}$) is defined on the $\sigma$-algebra $\mathcal{F}_\Theta \times \mathcal{F}_{j,\ell-2}^{k-1}(\boldsymbol{C})$. Similarly, if $\mu_i^{k-1}$ is defined on $\mathcal{F}_\Theta \times \{C_j^{k-2}, \emptyset\}$, then any probability measure $\mu_i^k$ in $\Delta(\Theta \times C_j^{k-1}, \mathscr{S}_i^k(\boldsymbol{C}))$ that satisfies the coherency condition is defined on $\mathcal{F}_\Theta \times \{C_j^{k-1}, \emptyset\}$. Finally, if $\mu_i^{k-1}$ is defined on the $\sigma$-algebra $\mathcal{F}_\Theta \times \mathcal{F}_{j,k-2}^{k-2}(\boldsymbol{C})$, then a probability measure $\mu_i^k$ in $\Delta(\Theta \times C_j^{k-1}, \mathscr{S}_i^k(\boldsymbol{C}))$ is coherent with $\mu_i^{k-1}$ only if it is defined on $\mathcal{F}_\Theta \times \mathcal{F}_{j,k-1}^{k-1}(\boldsymbol{C})$ or on $\mathcal{F}_\Theta \times \mathcal{F}_{j,k-2}^{k-1}(\boldsymbol{C})$. $\square$

## A.2 Proof of Lemma 5.1

It will be useful to introduce some notation and state some preliminary results. For any nonempty set $X$ and any nonempty collection $\mathcal{E}$ of subsets of $X$, let $\sigma(\mathcal{E})$ be the coarsest $\sigma$-algebra on $X$ that contains the sets in $\mathcal{E}$, that is, $\sigma(\mathcal{E})$ is the $\sigma$-algebra generated by $\mathcal{E}$.

The following preliminary result says that taking inverse images preserves $\sigma$-algebras:

**Lemma A.1.** Let $f : X \to Y$ be a function from $X$ into $Y$, and let $\mathcal{E}$ be a nonempty collection of subsets of $Y$. Then,

$$\sigma\big(\{f^{-1}(E) : E \in \mathcal{E}\}\big) = \{f^{-1}(E) : E \in \sigma(\mathcal{E})\}.$$

The proof is standard, and thus omitted. To state the second preliminary result, let $X$ be some nonempty set, and let $\mathscr{S}$ be a nonempty collection of $\sigma$-algebras on $X$. As before, $\Delta(X, \mathscr{S})$ is the collection of probability measures that are defined on some $\sigma$-algebra in $\mathscr{S}$. Let $\mathcal{A}$ be the family of sets of the form

$$\{\mu \in \Delta(X, \mathscr{S}) : \Sigma(\mu) = \mathcal{F}, \mu(E) \geq p\} : \qquad \mathcal{F} \in \mathscr{S}, E \in \mathcal{F}, p \in [0, 1],$$

and let $\mathcal{A}'$ be the family of sets of the form

$$\{\mu \in \Delta(X, \mathscr{S}) : E \in \Sigma(\mu), \mu(E) \geq p\} : \qquad \mathcal{F} \in \mathscr{S}, E \in \mathcal{F}, p \in [0, 1],$$

and let $\sigma(\mathcal{A})$ and $\sigma(\mathcal{A}')$ be the $\sigma$-algebras on $\Delta(X, \mathscr{S})$ generated by $\mathcal{A}$ and $\mathcal{A}'$, respectively. In general, these two $\sigma$-algebras can be different. However, as we show now, in an important class of cases, $\sigma(\mathcal{A})$ and $\sigma(\mathcal{A}')$ coincide:

**Lemma A.2.** Suppose $\mathscr{S}$ is countable and forms a filtration, and suppose there is $\underline{\mathcal{F}} \in \mathscr{S}$ such that $\underline{\mathcal{F}} \subseteq \mathcal{F}$ for all $\mathcal{F} \in \mathscr{S}$. Then $\sigma(\mathcal{A}) = \sigma(\mathcal{A}')$.

**Proof.** We first show that $\sigma(\mathcal{A}') \subseteq \sigma(\mathcal{A})$. It suffices to show that $\mathcal{A}' \subseteq \sigma(\mathcal{A})$. Fix $\mathcal{F} \in \mathscr{S}$, $E \in \mathcal{F}$, and $p \in [0, 1]$, and define

$$F' := \{\mu \in \Delta(X, \mathscr{S}) : E \in \Sigma(\mu), \mu(E) \geq p\},$$

so that $F' \in \mathcal{A}'$. It is immediate that $F' \in \sigma(\mathcal{A})$: Since for every $\mathcal{F}' \in \mathscr{S}$, either $E \in \mathcal{F}'$ or $E \notin \mathcal{F}'$, $F'$ is a countable union of sets in $\mathcal{A}$:

$$F' = \bigcup_{\mathcal{F}' \in \mathscr{S} : E \in \mathcal{F}'} \{\mu \in \Delta(X, \mathscr{S}) : \Sigma(\mu) = \mathcal{F}', \mu(E) \geq p\}.$$

Hence, $F' \in \sigma(\mathcal{A})$.

We next show that $\sigma(\mathcal{A}) \subseteq \sigma(\mathcal{A}')$. Again, fix $\mathcal{F} \in \mathscr{S}$, $E \in \mathcal{F}$, and $p \in [0, 1]$, and define

$$F := \{\mu \in \Delta(X, \mathscr{S}) : \Sigma(\mu) = \mathcal{F}, \mu(E) \geq p\},$$

so that $F \in \mathcal{A}$. If we show that $\Delta(X, \mathcal{F})$ is an element of $\sigma(\mathcal{A}')$, then we are done, because $F$ is then the intersection of two elements of $\sigma(\mathcal{A}')$:

$$F = \{\mu \in \Delta(X, \mathscr{S}) : E \in \Sigma(\mu), \mu(E) \geq p\} \cap \Delta(X, \mathcal{F}).$$

It remains to show that $\Delta(X, \mathcal{F}) \in \sigma(\mathcal{A}')$. Using that $\mathscr{S}$ is a countable filtration with a minimum element $\underline{\mathcal{F}}$, we can label the $\sigma$-algebras in $\mathscr{S}$ as

$$\underline{\mathcal{F}} =: \mathcal{F}_1 \subsetneq \mathcal{F}_2 \subsetneq \cdots$$

Then,

$$\Delta(X, \mathcal{F}_1) = \Delta(X, \mathscr{S}) \setminus \{\mu \in \Delta(X, \mathscr{S}) : E_2 \in \Sigma(\mu), \mu(E_2) \geq 0\}$$

for any $E_2 \in \mathcal{F}_2 \backslash \mathcal{F}_1$, so $\Delta(X, \mathcal{F}_1) \in \sigma(\mathcal{A}')$. For $k > 1$, assume that $\Delta(X, \mathcal{F}_1), \ldots, \Delta(X, \mathcal{F}_{k-1}) \in \sigma(\mathcal{A}')$. Then,

$$\Delta(X, \mathcal{F}_k) = \Delta(X, \mathscr{S}) \setminus \Big( \{\mu \in \Delta(X, \mathscr{S}) : E_{k+1} \in \Sigma(\mu), \mu(E_{k+1}) \geq 0\} \cup \\ \Delta(X, \mathcal{F}_1) \cup \cdots \cup \Delta(X, \mathcal{F}_{k-1}) \Big)$$

for any $E_{k+1} \in \mathcal{F}_{k+1} \setminus \mathcal{F}_k$, so $\Delta(X, \mathcal{F}_k) \in \sigma(\mathcal{A}')$. Since this holds for every $k$, and $\mathcal{F} = \mathcal{F}^k$ for some $k$, we have $\Delta(X, \mathcal{F}) \in \sigma(\mathcal{A}')$. $\square$

We can now prove Lemma 5.1. We will prove the result by induction. As part of the proof, we construct an inductive structure, in the following way. For each $k = 1, 2, \ldots$, we define a $\sigma$-algebra $\mathcal{Q}_i^k$ on $T_i$ for each player $i \in N$. (The $\sigma$-algebras $\mathcal{Q}_i^k$ will be the $\sigma$-algebras $\sigma(h_i^{\mathcal{T},k})$, defined below.) We then show that the $\sigma$-algebra of each type is either coincides with $\mathcal{Q}_i^m$ for some $m < k$, or is a superset of $\mathcal{Q}_i^k$. This gives us the inductive structure needed to prove the result. Note that Assumption 1 does not, by itself, provide such an order. In particular, it does not imply that $\mathcal{S}_i$ is a (countable) filtration.[26]

To prove the result, we define $\sigma(h_i^{\mathcal{T},1})$ to be the $\sigma$-algebra on $T_i$ that is generated by the function $h_i^{\mathcal{T},1}$, that is,

$$\sigma(h_i^{\mathcal{T},1}) := \{\{t_i \in T_i : h_i^{\mathcal{T},1}(t_i) \in B\} : B \in \mathcal{F}_{i,1}^1(\boldsymbol{C}^{\mathcal{T}})\}.$$

It will be notationally convenient to introduce the function $h_i^{\mathcal{T},0} : T_i \to \{x\}$, where $x$ is an arbitrary singleton, defined in the obvious way; thus, the $\sigma$-algebra $\sigma(h_i^{\mathcal{T},0})$ on $T_i$ generated by the function $h_i^{\mathcal{T},0}$ is simply the trivial $\sigma$-algebra $\{T_i, \emptyset\}$.

Lemmas A.3–A.5 help order the $\sigma$-algebras with which types are endowed, using the $\sigma$-algebras $\sigma(h_i^{\mathcal{T},0})$ and $\sigma(h_i^{\mathcal{T},1})$. Lemma A.3 is an auxiliary result that gives a useful characterization of $\sigma(h_i^{\mathcal{T},1})$.

**Lemma A.3.** The $\sigma$-algebra $\sigma(h_i^{\mathcal{T},1})$ is the coarsest $\sigma$-algebra on $T_i$ that dominates $\sigma(h_j^{\mathcal{T},0})$, i.e., $\sigma(h_i^{\mathcal{T},1}) \succ^* \sigma(h_j^{\mathcal{T},0})$.

[26] Indeed, it is possible to have $\mathcal{F}_i, \mathcal{F}'_i \in \mathcal{S}_i$ such that $\mathcal{F}_i \not\subseteq \mathcal{F}'_i$ and vice versa, or to have $\mathcal{F}_i^1, \mathcal{F}_i^2, \ldots, \mathcal{F}_i^{-1}, \mathcal{F}_i^{-2}, \ldots \in \mathcal{S}_i$ such that $\cdots \succ^* \mathcal{F}_i^{-1} \succ^* \mathcal{F}_i \succ^* \mathcal{F}_i^1 \succ^* \mathcal{F}_i^2 \succ^* \cdots$. It follows from the proof that for any such $\sigma$-algebra $\mathcal{F}_i$, we have $\mathcal{F}_i \supseteq \mathcal{Q}_i^m$ for all $m$, so that such $\sigma$-algebras do not affect the inductive structure.

**Proof.** Note that

$$
\begin{aligned}
\sigma\big(h_i^{\mathcal{T},1}\big) &= \Big\{\{t_i \in T_i : \mathrm{marg}_\Theta \beta_i(t_i) \in B\} \colon B \in \mathcal{F}^1_{i,1}(\boldsymbol{C}^{\mathcal{T}})\Big\} \\
&= \sigma\Big(\Big\{\{t_i \in T_i : \mathrm{marg}_\Theta \beta_i(t_i)(E) \geq p\} : E \in \mathcal{F}_\Theta, p \in [0,1]\Big\}\Big) \\
&= \sigma\Big(\Big\{\{t_i \in T_i : E' \in \mathcal{F}_\Theta \times \Sigma_i(t_i), \beta_i(t_i)(E') \geq p\} : E' \in \mathcal{F}_\Theta \times \sigma(h_j^{\mathcal{T},0}), p \in [0,1]\Big\}\Big),
\end{aligned}
$$

where the second equality uses Lemma A.1. □

**Lemma A.4.** For each type $t_i \in T_i$, we have $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},1})$ or $\Sigma_i(t_i) \supseteq \sigma(h_j^{\mathcal{T},1})$.

**Proof.** If $\Sigma_i(t_i) = \{T_j, \emptyset\}$, then clearly, $\Sigma_i(t_i) \subseteq \sigma(h_j^{\mathcal{T},1})$. If $\Sigma_i(t_i) \neq \{T_j, \emptyset\}$, then, by Assumption 1, there is $\mathcal{F}_i \in \mathcal{S}_i$ such that $\Sigma_i(t_i)$ dominates $\mathcal{F}_i$. Since any $\sigma$-algebra $\mathcal{F}_i \in \mathcal{S}_i$ is at least as fine as the trivial $\sigma$-algebra $\{T_i, \emptyset\}$, i.e., $\mathcal{F}_i \supseteq \{T_i, \emptyset\}$, $\Sigma_i(t_i)$ dominates $\{T_i, \emptyset\}$. But, by Lemma A.3, the $\sigma$-algebra $\sigma(h_j^{\mathcal{T},1})$ is the coarsest $\sigma$-algebra that dominates $\{T_i, \emptyset\}$. Hence, $\Sigma_i(t_i) \supseteq \sigma(h_j^{\mathcal{T},1})$. □

**Lemma A.5.** For each $t_i \in T_i$, if $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},1})$, then $\Sigma_i(t_i) = \sigma(h_j^{\mathcal{T},0})$.

**Proof.** Suppose $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},1})$. Then, by Assumption 1, either $\Sigma_i(t_i) = \{T_j, \emptyset\} = \sigma(h_j^{\mathcal{T},0})$, or there is a $\sigma$-algebra $\mathcal{F}_i \in \mathcal{S}_i$ such that $\Sigma_i(t_i)$ dominates $\mathcal{F}_i$. If there is such a $\sigma$-algebra $\mathcal{F}_i \in \mathcal{S}_i$, then an argument similar to the one in the proof of Lemma A.4 gives that $\Sigma_i(t_i) \supseteq \sigma(h_j^{\mathcal{T},1})$, a contradiction. □

For $k > 1$, assume inductively that for any $\ell \leq k-1$ and $i \in N$, the set $C_i^{\mathcal{T},\ell}$ has been defined and that the functions $h_i^{\mathcal{T},\ell}$ are well-defined. Let

$$
\sigma(h_i^{\mathcal{T},\ell}) = \{\{t_i \in T_i : h_i^{\mathcal{T},\ell}(t_i) \in B\} : B \in \mathcal{F}^\ell_{i,\ell}(\boldsymbol{C}^{\mathcal{T}})\}
$$

be the $\sigma$-algebra on $T_i$ that is generated by the function $h_i^{\mathcal{T},\ell}$. Also, assume that the following hold:

- the $\sigma$-algebra $\sigma(h_i^{\mathcal{T},\ell})$ is the coarsest $\sigma$-algebra on $T_i$ that dominates the $\sigma$-algebra $\sigma(h_j^{\mathcal{T},\ell-1})$;
- for each type $t_i \in T_i$, we have $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},\ell})$ or $\Sigma_i(t_i) \supseteq \sigma(h_j^{\mathcal{T},\ell})$;
- for each type $t_i \in T_i$, if $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},\ell})$, then there is $m < \ell$ such that $\Sigma_i(t_i) = \sigma(h_j^{\mathcal{T},m})$.

The next result shows that the function $h_i^{\mathcal{T},k}$ is well-defined:

**Lemma A.6.** For each type $t_i \in T_i$, we have $h_i^{\mathcal{T},k}(t_i) \in C_i^{\mathcal{T},k-1} \times \Delta(\Theta \times C_j^{\mathcal{T},k-1}, \mathscr{S}_i^k(\boldsymbol{C}^{\mathcal{T}}))$.

**Proof.** By the induction hypothesis, we have that the claim holds if and only if $\mu_i^k(t_i) = \beta_i(t_i) \circ (\mathrm{Id}_\Theta, h_j^{\mathcal{T},k-1})^{-1}$ is a probability measure in $\Delta(\Theta \times C_j^{\mathcal{T},k-1}, \mathscr{S}_i^k(\boldsymbol{C}^{\mathcal{T}}))$, where $\mathrm{Id}_\Theta$ is the identity function on $\Theta$. By the induction hypothesis, $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},k-1})$ or $\Sigma_i(t_i) \supseteq \sigma(h_j^{\mathcal{T},k-1})$. First suppose $\Sigma_i(t_i) \supseteq \sigma(h_j^{\mathcal{T},k-1})$. Then, for each $E \in \mathcal{F}_\Theta \times \mathcal{F}_{j,k-1}^{k-1}(\boldsymbol{C}^{\mathcal{T}})$, we have $(\mathrm{Id}_\Theta, h_j^{\mathcal{T},k-1})^{-1}(E) \in \mathcal{F}_\Theta \times \Sigma_i(t_i)$, so that $\mu_i^k(t_i)$ is a probability measure on $\mathcal{F}_\Theta \times \mathcal{F}_{j,k-1}^{k-1}(\boldsymbol{C}^{\mathcal{T}}) \in \mathscr{S}_i^k(\boldsymbol{C}^{\mathcal{T}})$. Next suppose that $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},k-1})$. By the induction hypothesis, there is $m < k-1$ such that $\Sigma_i(t_i) = \sigma(h_j^{\mathcal{T},m})$; let $m'$ be the largest $m' < k-1$ for which this holds. By a similar argument as before, it follows that $\mu_i^k(t_i)$ is a probability measure on $\mathcal{F}_\Theta \times \mathcal{F}_{j,m'}^{k-1}(\boldsymbol{C}^{\mathcal{T}})$, and this $\sigma$-algebra belongs to $\mathscr{S}_i^k(\boldsymbol{C}^{\mathcal{T}})$. $\square$

By Lemma A.6, we can define the $\sigma$-algebra

$$\sigma(h_i^{\mathcal{T},k}) := \{\{t_i \in T_i : h_i^{\mathcal{T},k}(t_i) \in B\} : B \in \mathcal{F}_{i,k}^k(\boldsymbol{C}^{\mathcal{T}})\}$$

on $T_i$ that is generated by the function $h_i^{\mathcal{T},k}$. We next establish the analogues of Lemmas A.3–A.5 for general $k$, to order the $\sigma$-algebras on the type sets.

**Lemma A.7.** The $\sigma$-algebra $\sigma(h_i^{\mathcal{T},k})$ is the coarsest $\sigma$-algebra on $T_i$ that dominates $\sigma(h_j^{\mathcal{T},k-1})$, i.e., $\sigma(h_i^{\mathcal{T},k}) \succ^* \sigma(h_j^{\mathcal{T},k-1})$.

**Proof.** By Lemma A.1, $\sigma(h_i^{\mathcal{T},k})$ is the coarsest $\sigma$-algebra that contains the sets in $\sigma(h_i^{\mathcal{T},k-1})$ as well as the sets

$$\{t_i \in T_i : \Sigma(\mu_i^k(t_i)) = \mathcal{F}, \mu_i^k(t_i)(E) \geq p\} \tag{A.1}$$

for $\mathcal{F} \in \mathscr{S}_i^k(\boldsymbol{C}^{\mathcal{T}})$, $E \in \mathcal{F}$, and $p \in [0,1]$. Since beliefs are coherent, that is, for all $\ell \leq k-1$,

$$\mathrm{marg}_{\Theta \times C_j^{\mathcal{T},\ell-1}} \mu_i^k(t_i) = \mu_i^\ell(t_i),$$

the $\sigma$-algebra $\sigma(h_i^{\mathcal{T},k})$ is the $\sigma$-algebra generated by the sets in (A.1). Since $\mathscr{S}_i^k(\boldsymbol{C}^{\mathcal{T}})$ is a countable filtration with a minimal element, it follows from Lemma A.2 that $\sigma(h_i^{\mathcal{T},k})$ is generated by the sets

$$\{t_i \in T_i : E \in \Sigma(\mu_i^k(t_i)), \mu_i^k(t_i)(E) \geq p\}$$

for $\mathcal{F} \in \mathscr{S}_i^k(\boldsymbol{C}^{\mathcal{T}})$, $E \in \mathcal{F}$, and $p \in [0,1]$. Using that for each $\mathcal{F} \in \mathscr{S}_j^k(\boldsymbol{C}^{\mathcal{T}})$, we have $\mathcal{F} \subseteq \mathcal{F}_\Theta \times \mathcal{F}_{j,k-1}^{k-1}(\boldsymbol{C}^{\mathcal{T}})$, we have that $\sigma(h_i^{\mathcal{T},k})$ is generated by the sets

$$\{t_i \in T_i : E \in \Sigma(\mu_i^k(t_i)), \mu_i^k(t_i)(E) \geq p\}$$

for $E \in \mathcal{F}_\Theta \times \mathcal{F}_{j,k-1}^{k-1}(\boldsymbol{C}^{\mathcal{T}})$, and $p \in [0,1]$, or, equivalently, the sets

$$\{t_i \in T_i : E' \in \mathcal{F}_\Theta \times \Sigma_i(t_i), \beta_i(t_i)(E') \geq p\}$$

for $E \in \mathcal{F}_\Theta \times \sigma(h_j^{\mathcal{T},k-1})$, and $p \in [0,1]$. Hence, $\sigma(h_i^{\mathcal{T},k}) \succ^* \sigma(h_j^{\mathcal{T},k-1})$. $\square$

**Lemma A.8.** For each $t_i \in T_i$, we have $\Sigma_i(t_i) \subseteq \sigma(h_j^{\mathcal{T},k})$ or $\Sigma_i(t_i) \supseteq \sigma(h_j^{\mathcal{T},k})$.

**Proof.** If $\mathcal{F}_i = \{T_i, \emptyset\}$, then clearly $\mathcal{F}_i \subseteq \sigma(h_i^{\mathcal{T},k})$. So suppose $\mathcal{F}_i \neq \{T_i, \emptyset\}$. By Assumption 1, one of the following holds:

(a) $\mathcal{F}_i$ is part of a mutual-dominance pair, that is, there is $\mathcal{F}_j \in \mathcal{S}_j$ such that $\mathcal{F}_i \succ \mathcal{F}_j$ and vice versa; or

(b) $\mathcal{F}_i$ is part of a finite chain, that is, there exist $m < \infty$ and (distinct) $\sigma$-algebras $\mathcal{F}_j^1, \mathcal{F}_j^3, \ldots, \mathcal{F}_j^m \in \mathcal{S}_j$ and $\mathcal{F}_i^2, \mathcal{F}_i^4 \ldots, \mathcal{F}_i^m \in \mathcal{S}_i$ such that

$$\mathcal{F}_i \succ^* \mathcal{F}_j^1 \succ^* \mathcal{F}_i^2 \succ^* \cdots \succ^* \mathcal{F}_j^m = \{T_j, \emptyset\}$$

if $m$ is odd, and

$$\mathcal{F}_i \succ^* \mathcal{F}_j^1 \succ^* \mathcal{F}_i^2 \succ^* \cdots \succ^* \mathcal{F}_i^m = \{T_i, \emptyset\}$$

if $m$ is even; or

(c) $\mathcal{F}_i$ is part of a cycle or infinite chain, that is, there exist $\sigma$-algebras $\mathcal{F}_j^1, \mathcal{F}_j^3, \ldots \in \mathcal{S}_j$ and $\mathcal{F}_i^2, \mathcal{F}_i^4, \ldots \in \mathcal{S}_i$ (where $\mathcal{F}_n^\ell, \mathcal{F}_n^m$ are not necessarily distinct, $n \in N$) such that

$$\mathcal{F}_i \succ^* \mathcal{F}_j^1 \succ^* \mathcal{F}_i^2 \succ^* \mathcal{F}_j^3 \succ^* \cdots .$$

We claim that if (a) or (c) is the case, then $\mathcal{F}_i \supseteq \sigma(h_i^{\mathcal{T},k})$. We present the argument for case (c); the argument for (a) is similar and thus omitted. Note that $\mathcal{F}_j^1 \supseteq \sigma(h_j^{\mathcal{T},0}) = \{T_j, \emptyset\}$. By the induction hypothesis, therefore, $\mathcal{F}_i \supseteq \sigma(h_i^{\mathcal{T},1})$. By a similar argument, $\mathcal{F}_j^1 \supseteq \sigma(h_j^{\mathcal{T},1})$. Since $\mathcal{F}_i$ dominates $\mathcal{F}_j^1$, it follows from the induction hypothesis and Lemma A.7 that $\mathcal{F}_i \supseteq \sigma(h_i^{\mathcal{T},2})$. Repeating this argument gives the desired result.

It remains to consider (b). We consider the case that $m$ is odd; the argument for the case that $m$ is even is similar. If $m \leq k$, then by the induction hypotheses and Lemma A.7, we have $\mathcal{F}_i = \sigma(h_i^{\mathcal{T},m}) \subseteq \sigma(h_i^{\mathcal{T},k})$. If $m > k$, then we have $\mathcal{F}_j^{m-k} = \sigma(h_j^{\mathcal{T},k})$ or $\mathcal{F}_i^{m-k} = \sigma(h_i^{\mathcal{T},k})$, depending on whether $k$ is odd or even. We treat the case that $\mathcal{F}_j^{m-k} = \sigma(h_j^{\mathcal{T},k})$; the argument for the case $\mathcal{F}_i^{m-k} = \sigma(h_i^{\mathcal{T},k})$ is similar. Since $\mathcal{F}_i^{m-k-1}$ dominates $\mathcal{F}_j^{m-k} \supseteq \sigma(h_j^{\mathcal{T},k-1})$, it follows from Lemma A.7 that $\mathcal{F}_i^{m-k-1} \supseteq \sigma(h_i^{\mathcal{T},k}) \supseteq \sigma(h_i^{\mathcal{T},k-1})$. By a similar argument, $\mathcal{F}_j^{m-k-2} \supseteq \sigma(h_j^{\mathcal{T},k}) \supseteq \sigma(h_j^{\mathcal{T},k-1})$. Repeating this argument gives $\mathcal{F}_i \supseteq \sigma(h_i^{\mathcal{T},k})$. $\square$

**Lemma A.9.** For each $t_i \in T_i$, if $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},k})$, then there is $m < k$ such that $\Sigma_i(t_i) = \sigma(h_j^{\mathcal{T},m})$.

**Proof.** Suppose $\Sigma_i(t_i) \subsetneq \sigma(h_j^{\mathcal{T},k})$. If $\mathcal{F}_i \subsetneq \sigma(h_i^{\mathcal{T},k-1})$, then the result follows from the induction hypothesis. So suppose $\mathcal{F}_i$ is not a strict subset of $\sigma(h_i^{\mathcal{T},k-1})$. By the induction hypothesis, we have $\mathcal{F}_i \supseteq \sigma(h_i^{\mathcal{T},k-1})$. If $\mathcal{F}_i = \sigma(h_i^{\mathcal{T},k-1})$, then we are done. So suppose $\mathcal{F}_i \supsetneq \sigma(h_i^{\mathcal{T},k-1})$. The proof is complete if we show that the joint hypothesis $\mathcal{F}_i \subsetneq \sigma(h_i^{\mathcal{T},k})$ and $\mathcal{F}_i \supsetneq \sigma(h_i^{\mathcal{T},k-1})$ leads to a contradiction. To derive a contradiction, use an argument similar to the one in the proof of Lemma A.8 to show that $\mathcal{F}_i \subsetneq \sigma(h_i^{\mathcal{T},k})$ implies that $\mathcal{F}_i$ is not part of a mutual-dominance pair, cycle or infinite chain. It then follows from Assumption 1 that $\mathcal{F}_i$ is part of a finite chain. But then $\mathcal{F}_i \supseteq \sigma(h_i^{\mathcal{T},k})$, or $\mathcal{F}_i = \sigma(h_i^{\mathcal{T},m})$ for some $m \leq k-1$, a contradiction. $\square$

This completes the induction. It follows that for each player $i \in N$ and $k = 1, 2, \ldots$, we have that $h_i^{\mathcal{T},k} : T_i \to C_i^{\mathcal{T},k-1} \times \Delta(\Theta \times C_j^{\mathcal{T},k-1}, \mathscr{S}_i^k(\boldsymbol{C}^{\mathcal{T}}))$ is well-defined. Also, note that for each $t_i \in T_i$, either $\Sigma_i(t_i) = \sigma(h_j^{\mathcal{T},k-1}) \subsetneq \sigma(h_j^{\mathcal{T},k})$, or $\Sigma_i(t_i) \supseteq \sigma(h_j^{\mathcal{T},m})$ for all $m$. $\square$

# Appendix B Proofs for Section 6

## B.1 Proof of Theorem 6.2 (continued)

We first complete the proof of the claim that if $\mathcal{T}^{\mathcal{H}}$ is not a Harsanyi extension of $\mathcal{T}^k$, then we can find a game and an equilibrium of the depth-$k$ model such that the corresponding strategy profile is not an equilibrium of the Harsanyi model, by considering the case in which $\beta_i^{\mathcal{H}}(t_i^{\mathcal{H}})(\theta) = 0$ (with $i$, $t_i^{\mathcal{H}}$, $t_i$, and $\theta$ as defined in the proof in the main text). Note that $\beta_i(t_i)(\theta) > 0$ by assumption, where we recall that $t_i = \varphi_i(t_i^{\mathcal{H}})$. Fix $z \geq (10 - 10\beta_i(t_i)(\theta))/\beta_i(t_i)(\theta)$, and consider the game $\mathcal{G}$ with action sets $S_i = \{s_i^1, s_i^2\}$ and $S_j = \{s_j\}$, and payoffs given by:

| $\theta$ | $s_j$ |
|---|---|
| $s_i^1$ | $z$,0 |
| $s_i^2$ | 0,0 |

| $\theta' \neq \theta$ | $s_j$ |
|---|---|
| $s_i^1$ | $-10$,0 |
| $s_i^2$ | 0,0 |

It is easy to see that the depth-$k$ model $(\mathcal{G}, \mathcal{T}^k)$ has an equilibrium in which $t_i$ plays $s_i^1$ with positive probability, while in any equilibrium of the Harsanyi model $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$, type $t_i^{\mathcal{H}}$ plays $s_i^2$ with probability 1. This concludes the proof that if $\mathcal{T}^{\mathcal{H}}$ is not a Harsanyi extension of $\mathcal{T}^k$, then there is a game and an equilibrium of the depth-$k$ model such that the corresponding strategy profile is not an equilibrium of the Harsanyi model.

We next prove the converse. Specifically, we show that for any game $\mathcal{G}$, and any Harsanyi extension $(\mathcal{T}^{\mathcal{H}}, \varphi)$ of $\mathcal{T}^k$, if $\sigma^k = (\sigma_i^k)_{i \in N}$ is an equilibrium of $(\mathcal{G}, \mathcal{T}^k)$, then $\sigma = (\sigma_i)_{i \in N}$, with

$\sigma = \sigma_i^k \circ \varphi_i$ for $i \in N$, is an equilibrium of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$. Fix a Harsanyi extension $(\mathcal{T}^{\mathcal{H}}, \varphi)$ of $\mathcal{T}^k$. Let $\mathcal{G}$ be a game, and suppose $\sigma^k$ is an equilibrium of $(\mathcal{G}, \mathcal{T}^k)$

Since $\sigma^k$ is an equilibrium of $(\mathcal{G}, \mathcal{T}^k)$, the strategy $\sigma_i^k$ is comprehensible for each type for $j$ in $\mathcal{T}^k$. It is straightforward to show that the strategy $\sigma_i^k \circ \varphi_i$ is comprehensible for each type for $j$ in $\mathcal{T}^{\mathcal{H}}$. Also, for each player $i \in N$, Harsanyi type $\tilde{t}_i^{\mathcal{H}} \in T_i^{\mathcal{H}}$, and action $s_i \in S_i$ such that $\sigma_i^k(\varphi_i(\tilde{t}_i^{\mathcal{H}}))(s_i) > 0$, we have that for all $b_i \in S_i$,

$$\begin{aligned}
\int_{\Theta \times T_j^{\mathcal{H}}} u_i(s_i, s_j, \theta) \sigma_j^k(\varphi_j(t_j^{\mathcal{H}}))(s_j) d\beta_i^{\mathcal{H}}(\tilde{t}_i^{\mathcal{H}}) &= \int_{\Theta \times T_j} u_i(s_i, s_j, \theta) \sigma_j^k(t_j)(s_j) d\beta_i(\tilde{t}_i) \\
&\geq \int_{\Theta \times T_j} u_i(b_i, s_j, \theta) \sigma_j^k(t_j)(s_j) d\beta_i(\tilde{t}_i) \\
&= \int_{\Theta \times T_j^{\mathcal{H}}} u_i(b_i, s_j, \theta) \sigma_j^k(\varphi_j(t_j^{\mathcal{H}}))(s_j) d\beta_i^{\mathcal{H}}(\tilde{t}_i^{\mathcal{H}}),
\end{aligned}$$

where $\tilde{t}_i := \varphi_i(\tilde{t}_i^{\mathcal{H}})$. The first and third lines use the standard change-of-variables result and that $\varphi$ is a type morphism, and the second line uses the best-response property. □

## B.2 Proof of Lemma 6.3

We start with an auxiliary result.

**Lemma B.1.** Let $\mathcal{T}^k$ be a depth-$k$ space that is $k$th-order nonredundant, and let $(\mathcal{T}^{\mathcal{H}}, \varphi)$ be a Harsanyi extension of $\mathcal{T}^k$. For each $i \in N$ and $t_i^{\mathcal{H}}, \tilde{t}_i^{\mathcal{H}} \in T_i^{\mathcal{H}}$, we have that $\varphi_i(t_i^{\mathcal{H}}) = \varphi_i(\tilde{t}_i^{\mathcal{H}})$ if and only if $h_i^{\mathcal{T}^{\mathcal{H}},k}(t_i^{\mathcal{H}}) = h_i^{\mathcal{T}^{\mathcal{H}},k}(\tilde{t}_i^{\mathcal{H}})$.

The proof is standard, and is included in the online appendix. We are now ready to prove Lemma 6.3. Let $\mathcal{T}^k$ be a depth-$k$ space that is $k$th-order nonredundant, and let $(\mathcal{T}^{\mathcal{H}}, \varphi)$ be a Harsanyi extension of $\mathcal{T}^k$. Suppose that for every game $\mathcal{G}$, for every Bayesian-Nash equilibrium $\sigma = (\sigma_i)_{i \in N}$ of $(\mathcal{G}, \mathcal{T}^{\mathcal{H}})$, there is a corresponding equilibrium of $(\mathcal{G}, \mathcal{T}^k)$. Fix a game $\mathcal{G}$, and let $\sigma$ be a Bayesian-Nash equilibrium of $(\mathcal{G},\mathcal{T}^{\mathcal{H}})$. Since there is a corresponding equilibrium of $(\mathcal{G}, \mathcal{T}^k)$, that is, the strategy profile $\sigma^k$ with $\sigma_i = \sigma_i^k \circ \varphi_i$ for $i \in N$ is an equilibrium of $(\mathcal{G}, \mathcal{T}^k)$, we have that for any $t_i^{\mathcal{H}}, \tilde{t}_i^{\mathcal{H}} \in T_i^{\mathcal{H}}$,

$$\varphi_i(t_i^{\mathcal{H}}) = \varphi_i(\tilde{t}_i^{\mathcal{H}}) \qquad \Longrightarrow \qquad \sigma_i(t_i^{\mathcal{H}}) = \sigma_i(\tilde{t}_i^{\mathcal{H}}).$$

By Lemma B.1, therefore, it follows that for any $t_i^{\mathcal{H}}, \tilde{t}_i^{\mathcal{H}} \in T_i^{\mathcal{H}}$,

$$h_i^{\mathcal{T}^{\mathcal{H}},k}(t_i^{\mathcal{H}}) = h_i^{\mathcal{T}^{\mathcal{H}},k}(\tilde{t}_i^{\mathcal{H}}) \qquad \Longrightarrow \qquad \sigma_i(t_i^{\mathcal{H}}) = \sigma_i(\tilde{t}_i^{\mathcal{H}}). \tag{B.1}$$

Since for each Harsanyi type space $\widetilde{\mathcal{T}}^{\mathcal{H}}$, there is a game $\widetilde{\mathcal{G}}$ and a Bayesian-Nash equilibrium $\widetilde{\sigma}$ of $(\widetilde{\mathcal{G}}, \widetilde{\mathcal{T}}^{\mathcal{H}})$ such that (B.1) does not hold if $h_i^{\mathcal{T}^{\mathcal{H}},m}(t_i^{\mathcal{H}}) \neq h_i^{\mathcal{T}^{\mathcal{H}},m}(\tilde{t}_i^{\mathcal{H}})$ for some $m \geq k$,[27] we

[27]For instance, take $\widetilde{\mathcal{G}}$ to be the game in the proof of Theorem 6.4.

must have that

$$h_i^{\mathcal{T}^{\mathcal{H}},k}(t_i^{\mathcal{H}}) = h_i^{\mathcal{T}^{\mathcal{H}},k}(\tilde{t}_i^{\mathcal{H}}) \qquad \Longrightarrow \qquad h_i^{\mathcal{T}^{\mathcal{H}},m}(t_i^{\mathcal{H}}) = h_i^{\mathcal{T}^{\mathcal{H}},m}(\tilde{t}_i^{\mathcal{H}})$$

for all $m \geq k$. That is, $\mathcal{T}^{\mathcal{H}}$ is an order-$k$ extension of $\mathcal{T}^k$. If $\mathcal{T}^{\mathcal{H}}$ is nonredundant, it is thus without loss of generality to take $T_i^{\mathcal{H}} = T_i$ for each $i \in N$. □

# References


Alaoui, L. and A. Penta (2013). Strategic thinking and the incentives to reason. Working paper, UW Madison and U Pompeu Fabra.

Aliprantis, C. D. and K. C. Border (2005). *Infinite Dimensional Analysis: A Hitchhiker's Guide* (3rd ed.). Berlin: Springer.

Battigalli, P. (1987). Comportemento razionale e d'equilibra nio giochi e nelle situazione sociali. Unpublished master thesis, Bocconi University.

Battigalli, P., A. Di Tillio, E. Grillo, and A. Penta (2011). Interactive epistemology and solution concepts for games with asymmetric information. *The B.E. Journal of Theoretical Economics 11*(1 (Advances)), Article 6.

Brandenburger, A. and E. Dekel (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory 59*, 189–198.

Brandenburger, A. and H. J. Keisler (2006). An impossibility theorem on beliefs in games. *Studia Logica 84*, 211–240.

Brocas, I., C. Camerer, J. Carrillo, and S. Wang (2009). Measuring attention and strategic behavior in games with private information. Working paper, Caltech.

Costa-Gomes, M., V. P. Crawford, and B. Broseta (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica 69*, 1193–1235.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2012). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*. Forthcoming.

Crawford, V. P. and N. Iriberri (2007). Level-$k$ auctions: Can a nonequilibrium model of strategic thinking explain the winners curse and overbidding in private-value auctions? *Econometrica 75*, 1721–1770.

Dekel, E., D. Fudenberg, and D. Levine (2004). Learning to play Bayesian games. *Games and Economic Behavior 46*, 282–303.

Dekel, E., D. Fudenberg, and S. Morris (2006). Topologies on types. *Theoretical Economics 1*, 275–309.

Dekel, E., D. Fudenberg, and S. Morris (2007). Interim correlated rationalizability. *Theoretical Economics 2*, 15–40.

Ely, J. and M. Peski (2011). Critical types. *Review of Economic Studies*. Forthcoming.

Ershov, M. P. (1974). Extension of measures and stochastic equations. *Theory of Probability and its Applications XIX*, 431–444.

Friedenberg, A. and H. J. Keisler (2011). Iterated dominance revisited. Working paper, Arizona State University.

Friedenberg, A. and M. Meier (2012). On the relationship between hierarchy and type morphisms. *Economic Theory 46*, 377–399.

Harsanyi, J. C. (1967–1968). Games on incomplete information played by Bayesian players. Parts I–III. *Management Science 14*, 159–182, 320–334, 486–502.

Heifetz, A. and W. Kets (2013). All types naive and canny. Working paper, Northwestern University.

Heifetz, A. and D. Samet (1998). Topology-free typology of beliefs. *Journal of Economic Theory 82*, 324–341.

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. (2005). Analogy-based expectation equilibrium. *Journal of Economic Theory 123*, 81–104.

Kechris, A. S. (1995). *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Berlin: Springer-Verlag.

Kets, W. (2009). Do you think about what I think you think? Finite belief hierarchies in games. Working paper, Stanford University.

Kets, W. (2010). Bounded reasoning and higher-order uncertainty. Working paper, Northwestern University.

Kets, W. (2013). Common belief with finite depth of reasoning. Working paper, Northwestern University.

Mertens, J.-F., S. Sorin, and S. Zamir (1994). Repeated games: Part A: Background material. Discussion Paper 9420, CORE.

Mertens, J. F. and S. Zamir (1985). Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory 14*, 1–29.

Morris, S., A. Postlewaite, and H. Shin (1995). Depth of knowledge and the effect of higher order uncertainty. *Economic Theory 6*, 453–467.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review 85*, 1313–1326.

Purves, R. (1966). Bimeasurable functions. *Fundamenta Mathematicae 58*, 149–157.

Qin, C. and C. Yang (2013). An explicit approach to modeling finite-order type spaces and applications. *Journal of Economic Theory*. Forthcoming.

Rogers, B. W., T. R. Palfrey, and C. F. Camerer (2009). Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory 144*, 1440–1467.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under "almost common knowledge". *American Economic Review 79*, 385–391.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons.

Stahl, D. O. and P. W. Wilson (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior 10*, 218–254.

Strzalecki, T. (2009). Depth of reasoning and higher order beliefs. Working paper, Harvard University.

Weinstein, J. and M. Yildiz (2007). A structure theorem for rationalizability with application to robust predictions of refinements. *Econometrica 75*, 365–400.