The Fundamental Theorem of Epistemic Game Theory:

The Infinite Case

EPICENTER Working Paper No. 25 (2021)

Stephan Jagau[AB*]

November 30, 2021

In auction theory, industrial organization, and other applications of games in economics, it is often convenient to let infinite strategy sets stand in for large finite strategy sets. A tacit assumption is that results for infinite games will translate back to their finite counterparts. Transfinite eliminations of non-best replies pose a radical challenge here, suggesting that *common belief in rationality* in infinite games may strictly refine *up to k-fold belief in rationality for all finite k*. I provide two equivalent characterizations of common belief in rationality for general purely measurable beliefs-type spaces. The first one is the usual transfinite elimination of non-best replies. The second one, elimination of non-best replies *and* supporting beliefs, entirely avoids transfinite eliminations. Hence, rather than revealing new depths of reasoning, transfinite eliminations signal an inadequacy of eliminating non-best replies as a general description for strategic rationality.

JEL classification: C72, D03, D83

Keywords: Common belief in rationality; Rationalizability; Epistemic game theory; Transfinite induction

Acknowledgments: I thank Christian Bach, Jean-Paul Carvalho, Stephanie Chan, Andrés Perea, Miklós Pintér, and Donald G. Saari for comments. Financial support from the Netherlands Organisation for Scientific Research (NWO) Grant 19.181SG.023 is gratefully acknowledged.

[A]IMBS, University of California, Irvine, Social Science Plaza A, Irvine, CA 92697-5100, USA
[B]EPICENTER, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands
[*]Email: sjagau@uci.edu. Web: https://sites.google.com/view/stephanjagau

One of the hallmarks of the epistemic approach to game theory is its distinction between modes of strategic reasoning on the one hand and procedural characterizations that select all strategies consistent with a certain reasoning on the other hand. The best-known result that spells out this distinction is the fundamental theorem of epistemic game theory (Brandenburger and Dekel 1987, Tan and da Costa Werlang 1988, Brandenburger 2014): For any finite $k \geq 1$, a strategy is consistent with rationality and up to k-fold belief in rationality iff it survives $k+1$-fold elimination of non-best replies (see Section 2 for definitions of these and all other concepts used here).

In many popular applications of game theory such as auction theory, bargaining theory, and industrial organization, the number k above can quickly grow very large, along with the vast number of (e.g.) bids, price levels, or production quantities that players may select from. Even more, one is naturally driven to consider *infinite* sets of strategies, given that the results of a game-theoretic analysis should ideally abstract from a specific grid of admissible alternatives (like, e.g., milligram production quantities, ¢-increments for prices and bids).

It is then intuitive that one would at least want to iterate elimination of non-best replies over the set of natural numbers $\omega = \{1, 2, 3, \dots\}$. And analogous to the finite case, one might hope that a strategy is consistent with up to k-fold belief in rationality for all finite $k \geq 1$ or *common* belief in rationality iff it survives ω-fold elimination of non-best replies.

Lipman (1994) was the first to show that this is generally *not* the case. Specifically, he develops a constructions for games with infinitely many strategies per player where one can keep eliminating non-best replies not only for ω rounds but for any countably ordinal number of rounds.[1]

How does this startling result relate to the fundamental theorem of epistemic game theory?

A natural idea would be that the correspondence between depths of reasoning and eliminations of non-best replies extends to the transfinite: I.e., for each infinite ordinal α, there would exist a reasoning depth of up to α-fold belief in rationality, such that a strategy is consistent with up to α-fold belief in rationality iff it survives $\alpha+1$-fold elimination of non-best replies.

While this resolution would nicely extend the structure of the fundamental theorem of epistemic game theory to all transfinite iterations of elimination of non-best replies, it would also spell trouble for the applications of infinite games mentioned above. With transfinite refinements $\alpha > \omega$ of up to α-fold belief in rationality, any infinite game model could now lead to predictions about rational strategic behavior that are specific to games with infinite strategy sets and, hence, inherently inapplicable to a finite reality.

In this paper, I show that strategic reasoning in infinite games never requires more of players than it does in all finite games – transfinite eliminations of non-best replies notwithstanding. To

[1] Beyond the initial infinite ordinal $\omega = \{1, 2, 3, \dots\}$, ordinal numbers such as $\omega + 1, \omega + 2, \dots$ correspond to different well-orderings of the natural numbers such as $2, 3, \dots, 1$ (order type $\omega + 1$), $3, 4, \dots, 1, 2$ (order type $\omega + 2$), $2, 4, \dots, 1, 3, \dots$ (order type 2ω). These order types measure the complexity of different ways in which we can arrange the elements in infinite sets like $\{1, 2, 3, \dots\}$, as opposed to order-independent magnitudes like $\aleph_0$, the cardinality of $\{1, 2, 3, \dots\}$. Formally, two sets are of equal cardinality iff there exists a bijection between them, whereas two *ordered* sets are of equal order type iff there exists a *monotonic* bijection between them.

achieve the maximum generality of results, I consider a belief-based model of reasoning in games using the purely measurable beliefs-type space originally constructed by Heifetz and Samet (1998b). As such, my modeling assumptions will coincide with what is needed to define a space of σ-additive hierarchies of probabilistic beliefs and to render strategic rationality a measurable event within that space. To my knowledge, this paper is the first to consider common belief in rationality from a purely measure-theoretic (rather than topological) point of view.[2]

Extending the fundamental theorem of epistemic game theory, I prove that all transfinite steps of elimination of non-best replies are *jointly* necessary and sufficient for a strategy to be rational given a theory of the game (i.e. a belief hierarchy) that expresses up to k-fold belief in rationality for all finite $k \geq 1$. So in particular, if a strategy survives transfinite step α but not $\alpha + 1$ of elimination of non-best replies, then, for *every* theory of the game (every belief hierarchy) that supports that strategy, there is a *finite* index k such that that theory is inconsistent with up to k-fold belief in rationality.

Hence transfinite steps of elimination of non-best replies do not imply new depths of strategic rationality that are specific to infinite game situations. Rather, what causes transfinite elimination steps in infinite games is the method of eliminating non-best replies by means of which common belief in rationality is typically characterized. To show this, I provide a fully general characterization of common belief in rationality via joint elimination of non-best replies *and* supporting beliefs. By picking up all constraints that increasing levels of up to k-fold belief in rationality impose on players' belief hierarchies, my characterization completely dispenses with transfinite eliminations.

The remainder of this paper is structured as follows: Section 1 revisits a well-known example of a game allowing for up to $\omega + 1$-fold elimination of non-best replies and argues that elimination steps ω and $\omega + 1$ cannot correspond to different states of mind of players in that game. Section 2 presents the general belief-based model of common belief in rationality and elimination of non-best replies and proves my core results. Section 3 uses a second example and a novel elimination procedure to show how transfinite eliminations are really caused by the *methods* we commonly apply to the study of common belief in rationality. Section 4 discusses related literature. Section 5 concludes. All proofs are in Appendix A.

1 An Introductory Example

A simple game where elimination of non-best replies can proceed for $\omega + 1$ steps is presented in Dufwenberg and Stegeman (2002), Example 3.

Example 1.1. ω+1 **Elimination Steps:** In a three-player game, players 1 and 2 choose quantities $q_i \in [0, 1]$, player 3 has a binary choice $q_3 \in \{E, N\}$. $b_i^1 \in \Delta(Q_{-i})$ denotes player i's first-order belief with $Q_{-i} = \times_{j \neq i} Q_j$ (see Section 2 for definitions). The expected utilities for players 1 and 2 are $U_i(q_i, b_i^1) = q_i \left(1 - q_i - \int_{[0,1]} q_j \, \mathrm{d}b_i^1(q_j)\right)$, $i, j \in \{1, 2\}$. So they engage in a standard Cournot

[2] See Section 4 for an overview of topological belief-based models of common belief in rationality and models of common knowledge of rationality.

competition, with Cournot-Equilibrium quantities at $q_i = \frac{1}{3},\ i = 1,2$. Utilities for player 3 are $U_3(E, b_3^1) = b_3^1\left(\left\{\frac{1}{3}, \frac{1}{3}\right\}\right)$ and $U_3(N, b_3^1) = 1 - b_3^1\left(\left\{\frac{1}{3}, \frac{1}{3}\right\}\right)$. Player 3, as it were, is taking a bet on whether the Cournot Equilibrium will be implemented or not: In the event that $q_i = \frac{1}{3},\ i = 1,2$, player 3 prefers E whereas they prefer N in all other cases.

It is easy to solve this game using iterated elimination of non-best replies, and it turns out that we need to eliminate $\omega + 1$ times: First, ω elimination steps are needed to select $q_i = \frac{1}{3},\ i = 1,2$ as the uniquely rationalizable strategies in the standard Cournot game between players 1 and 2. Then, given that players 1 and 2 will not select other quantities than $q_i = \frac{1}{3},\ i = 1,2$, we can perform an $\omega + 1$th elimination step since E is strictly better than N for player 3 whenever $b_3^1\left(\left\{\frac{1}{3}, \frac{1}{3}\right\}\right) = 1$.

I will now argue that there cannot exist belief hierarchies b_3 for player 3 that express ω-fold but not $\omega + 1$-fold belief in rationality. To show this, I solve Example 1.1 again. Different from before, I keep track of strategies *and* of first-order beliefs that support these strategies.

Step 1: All quantities $q_i > \frac{1}{2}$ are never a best reply for players 1 and 2. So these strategies and any supporting beliefs are eliminated at step 1. For all quantities $q_i \leq \frac{1}{2}$, the non-empty set of first-order beliefs such that q_i is a best reply is $B_i^1(q_i, 1) = \left\{ b_i^1 \in \Delta(Q_{-i}) \middle| \int_{[0,1]} q_j \, \mathrm{d}b_i^1(q_j) = 1 - 2q_i \right\},\ i, j \in \{1,2\}$. For player 3, E is a best reply given $b_3^1\left(\frac{1}{3}, \frac{1}{3}\right) \geq b_3^1\left([0,1]^2 \setminus \left\{\frac{1}{3}, \frac{1}{3}\right\}\right)$ and N is otherwise.

Step 2: Under 1-fold belief in rationality, players 1 and 2 put full measure on opponent quantities in $[0, \frac{1}{2}]$. So q_i is a best reply at step 2 iff the set $B_i^1(q_i, 2) = \left\{ b_i^1 \in \Delta(Q_{-i}) \middle| \int_{[0,\frac{1}{2}]} q_j \, \mathrm{d}b_i^1(q_j) = 1 - 2q_i \right\},\ i, j \in \{1,2\}$ is non-empty. Since the left-hand integral is at most $\frac{1}{2}$, only $\frac{1}{2} \geq q_i \geq \frac{1}{4}$ survive. All $q_i < \frac{1}{4}$ and any supporting beliefs are eliminated for players 1 and 2. For player 3 at step 1, $b_3^1\left(\left[0, \frac{1}{2}\right]^2\right) = 1$. Hence E is a best reply for player 3 iff $b_3^1\left(\frac{1}{3}, \frac{1}{3}\right) \geq b_3^1\left(\left[0, \frac{1}{2}\right]^2 \setminus \left\{\frac{1}{3}, \frac{1}{3}\right\}\right)$ and N is otherwise.

Step k: At step $k \geq 1$, players 1 and 2 put full measure on $Q(k-1) = \begin{cases} \left[\frac{1}{3}\left(1 - \frac{1}{2^k}\right), \frac{1}{3}\left(1 + \frac{1}{2^{k-1}}\right)\right], \text{ even } k \\ \left[\frac{1}{3}\left(1 - \frac{1}{2^{k-1}}\right), \frac{1}{3}\left(1 + \frac{1}{2^k}\right)\right], \text{ odd } k \end{cases}$.

So q_i is a best reply at step k iff $B_i^1(q_i, k) = \left\{ b_i^1 \in \Delta(Q_{-i}) \middle| \int_{Q(k-1)} q_j \, \mathrm{d}b_i^1(q_j) = 1 - 2q_i \right\},\ i, j \in \{1,2\}$ is non-empty. It follows that all $q_i \in Q(k)$ survive step k. All $q_i \in Q(k-1) \setminus Q(k)$ and any supporting belief are eliminated. For player 3 at step k, $b_3^1\left(q \in Q(k-1)^2\right) = 1$. Hence E is a best reply iff $b_3^1\left(\frac{1}{3}, \frac{1}{3}\right) \geq b_3^1\left(Q(k-1)^2 \setminus \left\{\frac{1}{3}, \frac{1}{3}\right\}\right)$ and N is otherwise.

Step ω: Combining all steps $k \geq 1$, players 1 and 2 put full measure on $\bigcap_{k \geq 0} Q(k) = \left[\frac{1}{3}, \frac{1}{3}\right]$. Hence $q_i = \frac{1}{3}$ is the unique best reply for players 1 and 2 after step ω. Thus far, everything is perfectly in line with what we found using strategy elimination without explicitly constructing supporting first-order beliefs. Different from before, however, E is the unique strategy for player 3 that survives ω-fold elimination of strategies and first-order beliefs. To see this, note that after combining all steps $k \geq 1$, we must have $b_3^1\left(\frac{1}{3}, \frac{1}{3}\right) = 1$. Hence, only strategy E survives step ω.

Eliminating strategies and first-order beliefs demonstrates that there exists no first-order belief supporting player 3's strategy N after step ω. Consequently, for every first-order belief b_3^1 support-

ing N, there is a finite index k such that b_3^1 does not survive step k. And hence, every supporting belief hierarchy b_3 can only ever rationalize strategy N up until some finite level k of up to k-fold belief in rationality. So at least in Example 1.1, elimination step ω is of a different nature than the finite steps $1, 2, \ldots$ and the ultimate step $\omega + 1$ in that no fixed depth of belief in rationality for player 3 can be associated with step ω. And furthermore, only those strategies that survive *all* transfinite steps of elimination of non-best replies can be supported by a belief hierarchy that expresses up to k-fold belief in rationality for all finite k.

In the following section, I show that *all* cases of games where transfinite elimination of non-best replies takes multiple steps have precisely this structure. Common belief in rationality is equivalent to up to k-fold belief in rationality for all finite k, but only transfinite elimination of non-best replies generally selects the strategies consistent with that reasoning depth.

2 A Belief-Based Analysis for Infinite Games

2.1 Measure-Theoretic Preliminaries

For any measurable space (X, Σ), $\Delta(X, \Sigma)$ will denote the set of σ-additive probability measures over X. For any countable product of sets $\times_k X_k$, I will consider the product σ-algebra. In case $X = \Delta(Y)$ is itself a space of probability measures on some measurable space (Y, Σ), I will consider the σ-algebra Σ_Δ generated by sets of the form $B^p(E) = \{\mu \in \Delta(Y, \Sigma) |\ \mu(E) \geq p\}$ for every $E \in \Sigma$ and $p \in [0, 1]$. At times, I will consider subsets $A \subseteq X$ that are non-measurable with respect to (X, Σ). In this case, I consider A as a measurable space equipped with the A-restriction of Σ, given by $\Sigma|A = \{F \subseteq A |\ F = E \cap A,\ E \in \Sigma\}$. Note that each $\mu_A \in \Delta(A, \Sigma|A)$ can be associated with a unique measure $\mu \in \Delta(X, \Sigma)$ via $\mu(E) = \mu_A(E \cap A)$ for $E \in \Sigma$.[3]

For ease of notation, I will write $\Delta(X)$ instead of $\Delta(X, \Sigma)$, whenever the σ-algebra over a given space X is understood.

2.2 Static Games

To start, I give a definition of static games. My definition is slightly different from standard ones. For all players i, I include not only players' strategies S_i and utility functions U_i but also all belief hierarchies B_i of players. As described earlier, each belief hierarchy b_i is a sequence of probability distributions $(b_i^1, b_i^2, \ldots)$, mapping a full theory that i could form about behavior and strategic reasoning in the game. For all of the analysis, I restrict to belief hierarchies satisfying coherency and common belief in coherency.[4] The construction of the set B_i proceeds in the spirit of Brandenburger and Dekel (1993). While their analysis assumes strategy sets S_i to be Polish,

[3] See Heifetz and Samet (1999), Lemma 2.2.

[4] Coherency requires that every belief hierarchy $(b_i^1, b_i^2, \ldots)$ satisfy $b_i^n = \text{marg}_{C_{-i} \times B_{-i}^{n-1}} b_i^{n+1}$, $n \geq 1$, where B_{-i}^{n-1} denotes the set of opponents' $n-1$th-order beliefs. Intuitively, within a fixed belief hierarchy, we can consistently reduce higher-order beliefs to lower-order beliefs through marginalization. Moreover, this is commonly believed. I.e., b_i^3 assigns full probability to opponents' coherent second-order beliefs b_j^2, $j \neq i$, b_i^4 assigns full probability to opponents' coherent third-order beliefs b_j^3, $j \neq i$ such that, for every opponent $j \neq i$, every b_j^3 assigns full probability to coherent opponents' second-order beliefs b_k^2, $k \neq j$, and so on.

results in Heifetz and Samet (1998b) imply a generalization to arbitrary sets.[5]

Following Heifetz and Samet (1998b), every belief hierarchy $b_i \in B_i$ is homeomorphic to a probability distribution over opponents' strategies and belief hierarchies $\delta(b_i) \in \Delta(S_{-i} \times B_{-i})$. As a corollary, for each set of nth-order beliefs $B_i^n = \text{proj}_{S_j \times B_j^{n-1}} B_i$, we find that every $b_i^n \in B_i^n$ is homeomorphic to a probability distribution over opponents' strategies and $n-1$th-order beliefs $\tilde{\delta}(b_i^n) \in \Delta(S_{-i} \times B_{-i}^{n-1})$. Therefore I identify b_i with $\delta(b_i)$ and b_i^n with $\tilde{\delta}(b_i^n)$ whenever that is useful.

Definition 2.1. *(Static Game)*
A static game is a tuple $\Gamma = (S_i, B_i, U_i)_{i \in I}$ *with* I *an arbitrary set of players,* S_i *a separable*[6] *set of strategies for player* i*,* B_i *the set of belief hierarchies for player* i *expressing coherency and common belief in coherency,*[5] *and* U_i *a measurable and bounded utility function* $U_i : S_i \times B_i^1 \to \mathbb{R}$ *with* $B_i^1 = \Delta(S_{-i})$ *and* $S_{-i} = \times_{j \neq i} S_j$*.*

Comments on Definition 2.1:

1. Usually definitions of static games do not include the set of belief hierarchies. Here, including belief hierarchies is useful since this equips every game with the full set of doxastic states that can be distinguished for each player. This provides us with all information that is needed to assess the relationship between elimination of non-best replies and (belief in) rationality.

2. I consider arbitrary non-expected utility functions $U_i : S_i \times B_i^1 \to \mathbb{R}$ for every player i. By contrast, the approach in previous literature (e.g. Arieli 2010) has been to start from integrable functions $u_i : S_i \times S_{-i} \to \mathbb{R}$, such that U_i has an expected utility form $U_i(s_i, b_i^1) = \int_{S_{-i}} u_i(s_i, s_{-i})\, \mathrm{d}b_i^1$. For finite games, imposing expected utility has a computationally attractive implication known as Pearce's Lemma (Pearce 1984): A strategy $s_i \in S_i$ for player $i \in I$ is a non-best reply iff it is strictly dominated by a randomized strategy. That is, there exists a measure $r \in \Delta(S_i)$ such that $u_i(s_i, s_{-i}) < \int_{S_i} u_i(s_i', s_{-i})\, \mathrm{d}r$ for all $s_{-i} \in S_{-i}$. For infinite games with non-expected utility as in Definition 2.1, a strategy being strictly dominated, clearly, is neither necessary nor sufficient for it to be a non-best reply. But even if we assume expected utility, strict dominance is merely sufficient, not necessary. Strategies can still be non-best

[5] If S_i is not Polish, hierarchies of coherent beliefs $b_i^1, b_i^2, \ldots$ need not induce distributions on opponents' strategies and coherent beliefs anymore (Heifetz and Samet 1999). Instead, $b_i^1, b_i^2, \ldots$ might distribute on a decreasing sequence of sets $X_i^n \subset S_{-i} \times B_{-i}^{n-1}$ such that no σ-additive extension on the space of belief hierarchies is possible. As shown in Heifetz and Samet (1998b), there still exists a universal type space, though, and we can then retroactively define the space B_i as the space of belief hierarchies encoded by these types. Alternatively, a more explicit remedy is to require the existence of a σ-additive extension in the definition of coherent belief hierarchies. As shown in Heifetz and Samet (1999), a space of collectively coherent belief hierarchies such that each $b_i \in B_i$ can be identified with a unique measure $\delta(b_i) \in \Delta(S_{-i} \times B_{-i})$ can be "carved out" from the space of collectively coherent hierarchies via transfinite induction. At each step of their transfinite induction, a new layer of "existence of σ-additive extensions and up to k-fold belief in σ-additive extensions" is added. This makes it apparent that the space of belief hierarchies encoded by the types in Heifetz and Samet's (1998b) universal type space is the natural extension of Brandenburger and Dekel (1993) to the purely measurable case, and that the structural properties of this space remain essentially unchanged in the general setting.

[6] Separability of the S_i will be necessary to ensure that the set of rational strategy-belief tuples is measurable, and (hence) that strategy belief-tuples expressing rationality and increasing orders of belief in rationality can be defined.

replies without being strictly dominated by any randomized strategy.[7] Given its broader scope in a belief-based setting, elimination of non-best replies is preferable for investigating how common belief in rationality relates to possibly transfinite procedural characterizations of behaviors consistent with that reasoning.[8]

2.3 Rationality and up to k-Fold Belief in Rationality

Here, I define rationality and up to k-fold belief in rationality and k-fold elimination of non-best replies. Replicating Brandenburger and Dekel (1987), I show that a strategy s_i can be rationalized under up to k-fold belief in rationality iff it survives $k+1$-fold elimination of non-best replies.

Definition 2.2. *(Rationality)*
Strategy $s_i \in S_i$ is rational for player i given belief hierarchy $b_i \in B_i$ if b_i induces a first-order belief $b_i^1 = \text{proj}_{S_{-i}} b_i$ such that $U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1),\ \forall s_i' \in S_i$.

I write $R_i(1) = \{(s_i, b_i) \in S_i \times B_i |\ U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \forall s_i' \in S_i\}$ for the set of (s_i, b_i)-tuples such that s_i is rational given b_i. In order to recursively define the sets of (s_i, b_i)-tuples such that b_i rationalizes s_i under up to k-fold belief in rationality, I first prove that the set $R_i(1)$ is measurable.

Lemma 2.3.
Let Γ be a static game. Then for any player i, the set $R_i(1)$ is measurable.

Proof. In Appendix. □

Going on, I define rationality and up to k-fold belief in rationality.

Definition 2.4. *(Rationality and up to k-Fold Belief in Rationality)*[9,10]
Recursively define

$$R_i(1) = \left\{(s_i, b_i) \in S_i \times B_i |\ U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \forall s_i' \in S_i\right\},$$

$$R_i(k) = R_i(1) \cap \left(S_i \times \Delta(R_{-i}(k-1))\right),\ k > 1.$$

For any $k \geq 0$, strategy s_i is rationalized by a belief hierarchy b_i under up to k-fold belief in rationality if $(s_i, b_i) \in R_i(k+1)$.

Oftentimes, we are not interested in figuring out exactly which belief hierarchy b_i for player i might rationalize a strategy s_i at some level of up to k-fold belief in rationality. Rather, we simply

[7] An example is in Arieli (2010). To bring Pearce's Lemma into the picture in infinite games, what one would need is expected utility *and* compact strategy sets S_i, $i \in I$. A proof can be provided upon request.

[8] For a knowledge-based setting without probabilistic beliefs, strategies that are undominated by any pure strategy would take the role of rational strategies in Definitions 2.2, 2.4, and 2.7 below. For that setting, Samet (2015) relates common knowledge of rationality to transfinite iterations of strict dominance, with results that are qualitatively similar to what I report in Observation 2.8 and Theorem 2.9 below.

[9] Given that $R_i(1)$ is measurable for any player i, it is straightforward to show that, for every player i and $k \geq 1$, $R_i(k)$ is a measurable set. The same goes for $R_i(\omega) = \bigcap_k R_i(k)$.

[10] Under the weak assumptions of Definition 2.1, rational strategies (Definition 2.2) might not exist for some or all beliefs b_i of a player i, potentially causing $R_i(k)$ to be empty starting at some finite $k \geq 1$. While this does not affect the results presented here, it does matter for how one needs to define elimination of non-best replies (Definition 2.5). See Milgrom and Roberts (1990), Dufwenberg and Stegeman (2002), Apt (2007), Chen et al. (2007) for details.

want to know which strategies s_i for player i are *consistent* with a given level of up to k-fold belief in rationality. The usual tool to answer this question is k-fold elimination of non-best replies.

Procedure 2.5. *(k-Fold Elimination of Non-Best Replies)*
Step 1*: For every player* $i \in I$*, let* $BR_i(1) = \{s_i \in S_i | \ \exists b_i^1 \in B_i^1 \ s.th. \ U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \forall s_i' \in S_i\}$.

Step $k > 1$*: Assume* $BR_i(k-1)$ *is defined for every player* i*. Then, for every player* i*,*

$$BR_i(k) = \{s_i \in S_i | \ \exists b_i^1 \in B_i^1 \ s.th. \ U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \forall s_i' \in S_i \ and \ \exists \mu \in \Delta(BR_{-i}(k-1))$$
$$s.th. \ b_i^1(E) = \mu(E \cap BR_{-i}(k-1)) \ for \ every \ measurable \ E \subseteq S_{-i}\}.$$

Comments on Procedure 2.5:

1. At every step of elimination of best replies, I reimpose the initial rationality constraint – making my procedure a generalization of the procedure from Bernheim (1984). This way of defining elimination of non-best replies yields a robust correspondence with increasing layers of rationality and belief in rationality – even where rational choices fail to exist for some beliefs. This is in line with previous results, see Milgrom and Roberts (1990), Apt (2007), Chen et al. (2007).

2. Invoking the measure μ in the definition of $BR_i(k)$, for $k \geq 1$ deals with the possibility that $BR_{-i}(k)$ might not be in the product σ-algebra Σ over S_{-i}.[11] While proceeding this way ensures that iterations of $BR_i(k)$ for $k \geq 1$ (including the further transfinite iterations in Definition 2.9) are well-defined, backing out a first-order belief from a measure over the $BR_{-i}(k)$-restriction of Σ might initially seem like a conceptually unappealing way of generalizing more standard versions of elimination of non-best replies. After all, invoking $BR_{-i}(k)$-restrictions whenever $BR_{-i}(k)$ is not measurable for some $k \geq 1$ does not alter the fact that Σ simply cannot talk about the (non)event that opponents choose strategies surviving k-fold elimination of non-best replies. However, as noted earlier, for any $k \geq 1$, the sets of opponents' strategy-belief hierarchy combinations $R_{-i}(k)$ are guaranteed to be measurable with respect to the product σ-algebra over $S_{-i} \times B_{-i}$. As such, non-measurable instances of best-reply sets do *not* indicate that players are unable to reason about strategically relevant aspects of a game – only that our technique for describing behavior consistent with such reasoning (elimination of non-best replies) does not necessarily assume any meaning within the theories (belief hierarchies) that players form in a given game.[12]

[11] Note that $BR_{-i}(k) \in \Sigma$ for all $k \geq 1$ implies $\mu \in \Delta(S_{-i}, \Sigma)$ so that Definition 2.5 reverts back to the standard definition of elimination of non-best replies whenever the best-reply sets are measurable with respect to the product σ-algebra over S_{-i}.

[12] In some sense, these observations only lend further support to the line of argumentation in this paper. In particular, Theorem 2.6, demonstrates that the non-measurability of the best-reply sets $BR_i(k)$, $k \geq 1$ for a given player i does not prevent them from being reliably associated with increasing iterations of the sets $R_i(k)$, $k \geq 1$. So, also in the fully general case, the sets $BR_i(k)$, $k \geq 1$ remain no less reliable indicators of which strategies of a player are compatible with a certain (finite) depth of strategic rationality. Theorem 2.10 takes this one step further by showing that only the final output of transfinite eliminations of non-best replies meaningfully relates to some depth of strategic reasoning. This demonstrates that, in the general case, the indication of what behaviors are

Using standard induction, I prove that, for any finite $k \geq 0$, strategy s_i survives $k+1$-fold elimination of non-best replies iff it is consistent with up to k-fold belief in rationality (consistency with 0-fold belief in rationality means that s_i is rational given some b_i).

Theorem 2.6. *(Consistency with up to k-Fold Belief in Rationality)*
Let Γ be a static game. For any player i, any strategy $s_i \in S_i$, and any finite $k \geq 0$, there exists a belief hierarchy $b_i \in B_i$ such that $(s_i, b_i) \in R_i(k+1)$ iff $s_i \in BR_i(k+1)$.

Proof. In Appendix. □

As we will see next, for ω-fold elimination of non-best replies and consistency with common belief in rationality, this one-to-one relationship generally fails to hold.

2.4 Common Belief in Rationality

Here, I define rationality and common belief in rationality, as well as ω-fold and transfinite elimination of non-best replies. As usual, I define common belief in rationality as the countable intersection of up to k-fold belief in rationality for all finite k, and I prove that this definition captures the maximal depth of strategic reasoning in all games. Next, I prove that a strategy s_i can be rationalized under common belief in rationality iff it survives *all* steps of transfinite elimination of non-best replies. Also, under additional compact-continuity type assumptions, I prove that surviving ω-fold elimination of non-best replies is necessary and sufficient.

Definition 2.7. *(Rationality and Common Belief in Rationality)*[9,13]
Assume $R_i(k)$ is defined for any finite $k \geq 1$. Then define $R_i(\omega) = \bigcap_{k<\omega} R_i(k)$. Strategy s_i is rationalized by a belief hierarchy b_i under common belief in rationality if $(s_i, b_i) \in R_i(\omega)$.

The following observation shows that common belief in rationality as in Definition 2.7 is *terminal* – any additional rationality constraint on $(s_i, b_i) \in R_i(\omega)$ is already accounted for by previously imposed constraints. This formally establishes that transfinite elimination of non-best replies cannot imply depths of reasoning beyond up to k-fold belief in rationality for all finite k.

For the purposes of the observation, let $R_i(\omega+1) = R_i(1) \cap (S_i \times \Delta(R_{-i}(\omega)))$.

Observation 2.8. *(Common Belief in Rationality is Terminal)*
Let $R_i(\omega+1) = R_i(\omega)$ for all players i.

Proof. In Appendix. □

Analogous to the finite case, strategies that are consistent with common belief in rationality can be studied using ω-fold and transfinite elimination of non-best replies.

compatible with certain depths of strategic reasoning also exhausts the usefulness of the non-best replies procedure. In particular, intermediate steps of transfinite elimination of non-best replies do not meaningfully relate to behaviors consistent with any particular depth of strategic rationality.

[13] As for Definition 2.4 above, the set $R_i(\omega)$ might be empty under the assumptions of Definition 2.1. Again, this does not impact my results. Regarding sufficient conditions for non-emptiness, first note that $R_i(k)$ is non-empty for every $k \geq 1$ under the compact-continuity conditions from Theorem 2.10 below. With Cantor's intersection theorem, it is then straightforward to show that also $R_i(\omega)$ will be non-empty in this case.

Procedure 2.9. *(ω-Fold and Transfinite Elimination of Non-Best Replies)*
Assume $BR_i(k)$ is defined for every player i and every $k \in \omega$.
Now, for every player i, let $BR_i(\omega) = \bigcap_{k\in\omega} BR_i(k)$.
Furthermore, for every successor ordinal $\alpha > \omega$, let

$$BR_i(\alpha) = \{s_i \in S_i |\ \exists b_i^1 \in B_i^1 \text{ s.th. } U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \forall s_i' \in S_i \text{ and } \exists \mu \in \Delta(BR_{-i}(\alpha-1))$$
$$\text{s.th. } b_i^1(E) = \mu(E \cap BR_{-i}(\alpha-1)) \text{ for every measurable } E \subseteq S_{-i}\}.$$

Lastly, for every limit ordinal $\alpha \geq \omega$, define $BR_i(\alpha) = \bigcap_{\beta<\alpha} BR_i(\beta)$.

I will say that $s_i \in S_i$ survives *transfinite* elimination of non-best replies if $s_i \in BR_i(\alpha)$ for every ordinal α.[14]

Even though common belief in rationality is equal to the countable intersection of up to k-fold belief in rationality for all finite $k \geq 1$, this does not mean that ω-fold elimination of non-best replies always selects strategies that are consistent with that depth of reasoning. This is shown in the next theorem.

In addition, I will provide conditions on strategy sets and utility functions for which ω rounds of elimination of non-best replies are indeed sufficient for consistency with common belief in rationality. In what follows, given a metric space of opponents' strategies S_{-i}, let $d(b_i^1, \hat{b}_i^1)$ denote the Lévy-Prokhorov distance between the probability distributions induced by first-order beliefs $b_i^1, \hat{b}_i^1 \in B_i^1$.

Theorem 2.10. *(Consistency with Common Belief in Rationality)*
Let Γ be a static game.

1. *For any player i and any strategy $s_i \in S_i$, there exists $b_i \in B_i$ such that $(s_i, b_i) \in R_i(\omega)$ iff s_i survives transfinite elimination of non-best replies.*
2. *For all players i, let S_i be compact and completely metrizable. Moreover, let U_i be such that, for every player i, every $s_i \in S_i$, every first-order belief b_i^1, and every $\varepsilon > 0$, there is a $\delta > 0$ such that for any belief $\hat{b}_i^1$ with $d(b_i^1, \hat{b}_i^1) < \delta$ we have $|U_i(c_i, b_i^1) - U_i(c_i, \hat{b}_i^1)| < \varepsilon$. Then s_i survives transfinite elimination of non-best replies iff $s_i \in BR_i(\omega)$.*

Proof. In Appendix. □

Comments on Procedure 2.5:

1. Theorem 2.10, Part 1 clarifies the doxastic meaning of transfinite eliminations of non-best replies: All transfinite elimination steps $\omega, \omega+1, \omega+2, \ldots$ *jointly* account for the requirements that common belief in rationality imposes on belief hierarchies.

[14] Since the set of *all* ordinals is a self-contradictory notion in Zermelo-Fraenkel set theory (Burali-Forti 1897), the only statement we can make about the length of the procedure is that it must terminate at *some* ordinal for any static game Γ. That the output of transfinite iterations converges at *some* ordinal wherever I use them is guaranteed by the well-ordering theorem in conjunction with the fact that $R_i(\omega)$ and $\text{proj}_{C_i} R_i(\omega)$ are best-response sets. Sufficient conditions for transfinite elimination of non-best replies to be of countably transfinite length $\alpha < \omega_1$ are given in Arieli (2010).

2. It is well-established that compact-continuity type assumptions yield a characterization of common belief in rationality through ω-fold elimination of non-best replies. Theorem 2.10, Part 2 proves a version of this result with compact and separable, completely metrizable (Polish) strategy sets, and with utility functions that are continuous in first-order beliefs with respect to the weak topology. This slightly generalizes the canonical result from Tan and da Costa Werlang (1988) who additionally assume expected utility.

3. Like the sufficiency-part of Theorem 2.6, the sufficiency-part of Theorem 2.10, Part 1 is constructive. That is, given an arbitrary strategy surviving transfinite elimination of non-best replies, I show how to *construct* a supporting belief hierarchy expressing common belief in rationality. Contrast this with the sufficiency-part of Theorem 2.10, Part 2, which needs to be proven using standard topological arguments around Cantor's intersection theorem.

3 What Causes Transfinite Eliminations?

Theorem 2.10 suggests that transfinite rounds of eliminations have less to do with the reasoning embodied by common belief in rationality than with the method of elimination of non-best replies that we commonly study it with. Here, I revisit an example from Lipman (1994) to zoom in on what causes transfinite eliminations.

Example 3.1. 2ω **Elimination Steps:** In a symmetric two-player game, strategy sets for players $i = 1, 2$ each consist of two countable components $A = \{a_1, a_2, \dots\}$ and $D = \{d_1, d_2, \dots\}$. Utilities for both players are of expected-utility form, so we can summarize them with the following matrix:[15]

	d_2	d_3	d_4	$\dots$	d_1	a_2	a_3	$\dots$	a_1
d_2	0	$-\frac{5}{2}$	$-\frac{5}{2}$	$\dots$	$-\frac{5}{2}$	$-\frac{3}{2}$	$-\frac{3}{2}$	$\dots$	$-\frac{3}{2}$
d_3	$\frac{7}{3}$	0	$-\frac{7}{3}$	$\dots$	$-\frac{7}{3}$	$-\frac{4}{3}$	$-\frac{4}{3}$	$\dots$	$-\frac{4}{3}$
d_4	$\frac{9}{4}$	$\frac{9}{4}$	0	$\dots$	$-\frac{9}{4}$	$-\frac{5}{4}$	$-\frac{5}{4}$	$\dots$	$-\frac{5}{4}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
d_1	2	2	2	$\dots$	0	-2	-2	$\dots$	-2
a_2	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\dots$	$\frac{3}{2}$	0	$-\frac{3}{2}$	$\dots$	$-\frac{3}{2}$
a_3	$\frac{4}{3}$	$\frac{4}{3}$	$\frac{4}{3}$	$\dots$	$\frac{4}{3}$	$\frac{4}{3}$	0	$\dots$	$-\frac{4}{3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_1	1	1	1	$\dots$	1	1	1	$\dots$	0

I will analyze this game two times – first using 2ω steps of elimination of non-best replies, then using ω-fold elimination of (s_i, b_i)-tuples so as to directly implement Definition 2.7.

[15] Note that, since $A \cup D$ is unbounded, this game is an example of a non-compact belief-continuous game.

Elimination of Non-Best Replies:

Step 1: It is easy to see that d_2 is always a worse reply than d_3. All other strategies are a best reply given some first-order belief for both players: First, we can check that d_k is the best reply to d_{k-1} for any $k > 2$. Similarly, a_k is the best reply to any a_{k-1} for $k > 2$, a_2 is the best reply to d_1, and a_1 is the best reply to itself. To show that d_1 is a best reply, we construct a supporting belief $b_i^1 \in \Delta(D \backslash \{d_1\})$. Note that d_1 yields 2 for such a belief. The strongest competitors for d_1 are the other strategies in D, so we need to make sure that, for all $k \geq 2$, $U_i(d_1, b_i^1) \geq U_i(d_k, b_i^1) \Leftrightarrow$ $2 \geq \left(2 + \frac{1}{k}\right)\left[b_i^1(c_j \in \{d_2, \ldots, d_{k-1}\}) - b_i^1(c_j \in \{d_{k+1}, \ldots\})\right]$. A sufficient condition follows from dropping the right-hand negative term: $\frac{2k}{2k+1} \geq b_i^1(c_j \in \{d_2, \ldots, d_{k-1}\})$. For $k = 2$ this is trivially satisfied. And since the expression is positive and converges to 1 as $k \to \infty$, we can construct beliefs that satisfy the requirement for all $k \geq 3$.

Step k: After eliminating d_{k-1} from players' first-order beliefs, d_k is strictly dominated. All other strategies that survived step $k-1$ can be shown to be best replies with the previous methods. Note that the set of supporting beliefs for d_1 shrinks towards the empty set as we iterate over finite k.

Step ω: At Step ω, we have eliminated all d_k, $k \geq 2$. All other strategies survive ω-fold elimination of non-best replies. However, we will show that only a_1 survives all transfinite elimination steps.

Step $\omega + 1$: At step $\omega + 1$, we find that d_1 is never a best reply since any supporting belief must put positive measure on the set $\{d_2, d_3, \ldots\}$. All strategies in A survive step $\omega + 1$.

Step $\omega + 2$: After eliminating d_1, a_2 is never a best reply. All $s_i \in A \backslash \{a_2\}$ survive step $\omega + 2$.

Step $\omega + k$, $k > 2$: After eliminating a_{k-1}, a_k is never a best reply. All $s_i \in A \backslash \{a_2, \ldots, a_k\}$ survive.

Step 2ω: At step 2ω, we have eliminated all $a_2, a_3, \ldots$. Hence a_1 is the unique best reply.

Elimination of (s_i, b_i)-Tuples:

By Theorem 2.10, all steps $\omega, \ldots, 2\omega$ must be crammed into a single step if we eliminate among (s_i, b_i)-tuples. Here is how this feat is achieved in practice:

Steps 1 through k: Since utility depends on strategies and first-order beliefs, I simplify the analysis by tracking (s_i, b_i^1)-tuples at step 1, (s_i, b_i^2)-tuples at step 2, and so on. At step 3, e.g., consider $C_i \times B_i^3$ and eliminate (1) all surviving $\{d_4\} \times B_i^3$ (rationality constraint), (2) all surviving (c_i, b_i^3) such that $\{d_3\} \times B_j^2 \in \text{Supp}(b_i^3)$ (1-fold belief in rationality constraint), and (3) all surviving (c_i, b_i^3) such that there is $(c_j, b_j^2) \in \text{Supp}(b_i^3)$ with $\{d_2\} \times B_i^1 \in \text{Supp}(b_j^2)$ (2-fold belief in rationality constraint). This captures all constraints that increasing levels of rationality put on (s_i, b_i)-tuples.

Combining the Constraints, Step ω: Although the additional constraints above are cumbersome and irrelevant for finite levels of up to k-fold belief in rationality, they come back to haunt us at the limit of common belief in rationality. To see this, consider what constraints are imposed as we combine the restrictions from up to k-fold belief in rationality for all finite $k \geq 0$:

0) All (s_i, b_i) with $s_i \in \{d_2, d_3, \ldots\}$ are eliminated ($\hat{=}$ ω-fold elimination of non-best replies).

1) All (s_i, b_i) with first-order beliefs b_i^1 that deem $\{d_2, d_3, \ldots\}$ possible are eliminated. This includes all (s_i, b_i) involving strategy d_1 ($\hat{=}$ $\omega + 1$-fold elimination of non-best replies).

2) All (s_i, b_i) with second-order belief b_i^2 that entertains a b_j^1 that deems $\{d_2, d_3, \dots\}$ possible are eliminated. Since any (s_i, b_i) with a first-order belief b_i^1 that deems d_1 possible for the opponent is among those, all such (s_i, b_i) are eliminated. Since any (s_i, b_i) rendering a_2 a best reply for player i involves a first-order belief b_i^1 that deems d_1 possible for the opponent, all (s_i, b_i) involving strategy a_2 are eliminated ($\hat{=}\, \omega + 2$-fold elimination of non-best replies).

k) All (s_i, b_i) with a kth-order belief b_i^k that deems $\{d_2, d_3, \dots\}$ possible at level k are discarded. Folding back, all (s_i, b_i) involving strategy a_k are eliminated ($\hat{=}\, \omega + k$-fold elim. of non-best replies).

ω) Continuing this reasoning for all levels of the belief hierarchy, we find that all belief hierarchies b_i that rationalize $a_2, a_3, \dots$ must deem $\{d_2, d_3, \dots\}$ possible at *some level* $k \geq 1$ of the belief hierarchy. Hence, in particular, all (s_i, b_i) involving strategies $a_2, a_3, \dots$ are eliminated at step ω of eliminating (s_i, b_i)-tuples. And, in fact, all belief hierarchies that even deem strategies other than a_1 possible must be eliminated following this reasoning. It follows that (a_1, b_i) where b_i puts full measure on a_1 at all levels of higher-order beliefs is the unique (s_i, b_i)-tuple expressing rationality and common belief in rationality for both players.

Example 3.1 makes the notion that ω-fold elimination of (s_i, b_i)-tuples (Definition 2.7) selects the same strategies as *any* transfinite number of eliminations of non-best replies (Procedure 2.9) more tangible: In the example, higher-order beliefs matter for characterizing common belief in rationality even though utilities depend only on strategies and first-order beliefs. This is caused by infinitely many new requirements on infinitely many levels of the belief hierarchy that are combined at the limit of common belief in rationality. As these restrictions interact, they fold back to eliminate additional strategies without requiring further elimination steps beyond step ω – if only we keep track of *all* rationality constraints that are imposed on each (s_i, b_i) by Definition 2.7.

Consequently, transfinite elimination steps mirror a secondary ordering of strategies, according to levels of beliefs that elimination procedures must track so as to correctly determine whether strategies are (in)consistent with common belief in rationality. E.g., in Example 3.1, ω-fold elimination of (s_i, b_i^1)-tuples eliminates all strategies other than $\{a_1, a_2, \dots\}$, ω-fold elimination of (s_i, b_i^2)-tuples eliminates all strategies other than $\{a_1, a_3, \dots\}$, etc.

Taking this logic a step further, for every game Γ and every player i in it, elimination of non-best replies imposes a unique order type $\mathcal{O}_\Gamma(S_i)$ on i's strategy set S_i, corresponding to the infinite ordinal at which elimination of non-best replies converges for that player. For any finite $m \geq 0$, $\mathcal{O}_\Gamma(S_i) = \omega + m$ then implies that ω-fold elimination of (s_i, b_i^m)-tuples selects all strategies in S_i that are consistent with common belief in rationality. This is the case in Example 1.1 where $\mathcal{O}_\Gamma(Q_1) = \mathcal{O}_\Gamma(Q_2) = \omega$ and $\mathcal{O}_\Gamma(Q_3) = \omega + 1$. Here, as previously observed, ω-fold elimination of strategies works for players 1 and 2 whereas ω-fold elimination of (s_i, b_i^1)-tuples does the trick for player 3. Similarly, for any ordinal $\alpha \geq 2\omega$, $\mathcal{O}_\Gamma(S_i) = \alpha$ implies that ω-fold elimination of (s_i, b_i)-tuples is needed to select the strategies in S_i that are consistent with common belief in rationality.

This is the case in Example 3.1 where $\mathcal{O}_\Gamma(S_i) = 2\omega$ for players $i = 1, 2$.[16]

So rather than revealing new depths of reasoning of players in infinite games, transfinite elimination steps signal an inadequacy of elimination of non-best replies as a general description for common belief in rationality. Richer procedures, using elimination of strategy-belief-tuples, make transfinite elimination steps unnecessary. To formally show that common belief in rationality can be characterized in a way that completely avoids transfinite eliminations, I present a procedure for general games such that any output survives ω steps of elimination iff it is consistent with common belief in rationality. Example 3.1 already suggests what might be the sparsest procedure of this kind. As we saw there, the fact that utility depends on (s_i, b_i^1)-tuples implies that the constraints from up to k-fold belief in rationality for any finite $k \geq 1$ must operate on (s_i, b_i^{k+1})-tuples. This is true more generally, leading to the following procedure and theorem.

Procedure 3.2. *(Elimination of Non-Best Replies and Supporting Beliefs)*[17]

Step 1*: For every player* $i \in I$*, define* $BR_i^*(1) = \{(s_i, b_i^1) \in S_i \times B_i^1 |\ u_i(s_i, b_i^1) \geq u_i(s_i', b_i^1),\ \forall s_i' \in S_i\}$.

Step $k > 1$*: Assume* $BR_i^*(k-1)$ *is defined for every player* i*. Then, for every player* i*,* $BR_i^*(k) = \{(s_i, b_i^k) \in S_i \times B_i^k |\ (s_i, b_i^{k-1}) \in BR_i^*(k-1),\ b_i^k \in \Delta(BR_{-i}^*(k-1))\}$.

Step ω*: For every player* i *and finite* $k \geq 1$*, let* $\overline{BR}_i^*(k) = \{(s_i, b_i) \in S_i \times B_i |\ (s_i, b_i^k) \in BR_i^*(k)\}$. *Now define* $BR_i^*(\omega) = \bigcap_{k<\omega} \overline{BR}_i^*(k)$.

Theorem 3.3. *(Exact Characterization of Common Belief in Rationality)*
Let Γ *be a static game. For any player* i *and any strategy* $s_i \in S_i$*, there exists a belief hierarchy* b_i *that rationalizes* s_i *under common belief in rationality iff* $s_i \in \text{proj}_{S_i} BR_i^*(\omega)$.

Proof. In Appendix. □

[16] Continuing this reasoning, one may ask what more additional information starts mattering for common belief in rationality as we move beyond 2ω steps of elimination. After all, elimination of (s_i, b_i)-tuples is already necessary to characterize common belief in rationality whenever $\mathcal{O}_\Gamma(S_i) = 2\omega$ for some player i. To get an idea, consider the reasoning I used when combining the finite steps of (s_i, b_i)-elimination in Lipman's (1994) game above. Effectively, what I could do in that game was to treat the outcome of ω-fold elimination of non-best replies (elimination of all $d_2, d_3, \dots$) as a belief restriction on all levels of a players' belief hierarchy. Additional eliminations, corresponding to steps $\omega + 1, \dots, 2\omega$ of elimination of non-best replies, then followed from this belief restriction combined with consistency with up to k-fold belief in rationality for every finite order of belief b_i^k, $k \geq 1$.

Clearly, if there were a step $2\omega + 1$ of elimination of non-best replies, this simplification would no longer be possible. Instead, one would now need to additionally explore the implications of restricting first-order beliefs to the outcome of not only ω-fold but 2ω-fold elimination of non-best replies to capture all constraints that common belief in rationality puts on players' belief hierarchies. More generally, for any $m, k \geq 1$, $\mathcal{O}_\Gamma(S_i) = m\omega + k$ for some player i would imply that the behaviorally relevant restrictions from common belief in rationality on i's belief hierarchy would follow from

1) treating the output of $(m-1)\omega$-fold elimination of non-best replies as a belief restriction on all levels of i's belief hierarchy,

2) treating the output of $m\omega$-fold elimination of non-best replies as a belief restriction on all of i's finite-order beliefs up to and including order k,

3) imposing consistency with up to n-fold belief in rationality on the so-restricted nth-order beliefs of player i for all $n \geq 1$.

Analogous observations could be made for even larger ordinals.

[17] It is easy to prove, by induction, that each set $BR_i^*(k)$, $k \geq 1$ is measurable with respect to the product σ-algebra over $S_i \times B_i^k$. The induction start is analogous to Lemma 2.3, using that U_i only depends on $S_i \times B_i^1$.

Comments on Theorem 3.3:

1. Theorem 3.3 formally shows that transfinite elimination steps can be completely avoided as soon as we use an elimination procedure that picks up all constraints that increasing levels of up to k-fold belief in rationality could possibly impose on players' belief hierarchies.[18]

2. One might feel that procedure 3.3 involves a big complexity jump coming from elimination of non-best replies (Procedure 2.5). What about games where ω-fold elimination of (s_i, b_i^n)-tuples for some finite $n \geq 1$ characterizes common belief in rationality? Two observations suggest that procedures of intermediate complexity would not be useful in practice. Firstly, these intermediate procedures would only ever work for games such that $\sup_{i \in I} \mathcal{O}_\Gamma(S_i) < 2\omega$. To see this, note that already in Example 3.1, the minimal information we would need to characterize behavior consistent with common belief in rationality is not furnished by (s_i, b_i^n)-tuples of any finite order n. Secondly, and more importantly, elimination of (s_i, b_i^n)-tuples for some finite $n \geq 1$ would only ever be useful if we could characterize the subset of games on which it works *a priori*. However, since the numbers $\mathcal{O}_\Gamma(S_i)$, $i \in I$ follow from performing transfinite elimination of non-best replies (Procedure 2.5) on Γ, knowing a set of games G for which some procedure of intermediate complexity does apply means that one has already solved the games in G using Procedure 2.9.

3. As we saw leading up to Theorem 3.3, the dependence of utilities u_i on (s_i, b_i^1) ultimately determines what information furnishes an exact characterization of common belief in rationality in the static games from Definition 2.1. We could replace standard utilities with bounded and measurable *psychological utility functions* $u_i : C_i \times B_i \to \mathbb{R}$ for all $i \in I$, and that would give us a general version of static *psychological* games. Remarkably, the results presented in this paper fully generalize to static psychological games in a rather straightforward way. Details are in Appendix B.

4 Related Literature

4.1 Common Belief in Rationality in Topological Spaces

In this paper, I have investigated the relationship between common belief in rationality and procedural characterizations of behavior that is consistent with common belief in rationality. Throughout I have considered the case of a purely measurable beliefs-type space (Heifetz and Samet 1998b) with separable strategy sets.

Arieli (2010) has previously provided an analysis of elimination of non-best replies for the case of Polish strategy- and beliefs-type-spaces (Brandenburger and Dekel 1993) with integrable utility

[18] As such, Procedure 3.2 also demonstrates that the "iterated mutual belief"-approach (characterizing common belief in rationality as the countable intersection of constraints from up to k-fold belief in rationality for all finite k) can be implemented at just the same level of generality as the "fixed point"-approach (characterizing common belief in rationality as the fixed point of a sequence of sets collecting constraints from increasing orders of up to k-fold belief in rationality, Aumann 1976) – at least in the case of the purely measurable beliefs-type space. In fact, it is not hard to see that Procedure 3.2 represents the sparsest implementation of the "iterated mutual belief"-approach that is still equivalent to the fixed point approach.

functions $u_i : S_i \times S_{-i} \to \mathbb{R}$ for every player i. Like in my setting, common belief in rationality itself admits no further refinements, but only surviving transfinite elimination of non-best replies is necessary and sufficient for a strategy to be consistent with common belief in rationality. Since the setting with Polish strategy- and type-spaces and integrable utility functions on $S_i \times S_{-i}$ is a special case of the purely measurable beliefs-type space with separable strategy sets and non-expected utility functions on $U_i : S_i \times B_i^1 \to \mathbb{R}$ for every player i, Arieli's (2010) characterization result follows from Theorem 2.10 above.

However, not only are the proof-techniques in both papers different, but also Arieli's (2010) setup shares many properties with finite games that break down in the purely measurable case. Most importantly, Arieli (2010) shows that all iterations of elimination of non-best replies are (Lebesgue) measurable. Hence, different from what was observed for my Definitions 2.5 and 2.9 above, a player can always express the event that opponents only choose strategies surviving a certain (finite or transfinite) number of eliminations of non-best replies within the same probabilistic language as the event that opponents exhibit rationality and increasing depths of belief in rationality.

In this sense, the results I have presented here help separate familiar properties of common belief in rationality and transfinite elimination of non-best replies that are driven by topological assumptions from other (more fundamental) characteristics that are invariant to specific assumptions regarding strategy sets, beliefs, and utilities. In particular, the purely measurable case reveals a clear distinction between strategic reasoning (Definitions 2.4 and 2.7) and behavioral evidence of strategic reasoning (Definitions 2.5 and 2.9), with the behavioral evidence being generally extraneous to the reasoning itself (and hence not necessarily measurable relative to the product σ-algebra over $S_i \times B_i$ for any player $i \in I$). A measure-theoretic approach to infinite games is uniquely suited to provide such general conclusions since a universal beliefs-type space fails to exist in the general topological case (see Pinter 2010).

4.2 Common Knowledge of Rationality

A number of papers have studied the properties of transfinite elimination of non-best replies (or transfinite iterations of strict dominance) in relation to common knowledge of rationality. Examples include Lipman (1994) and Samet (2015). Notably, these papers report results that are qualitatively similar to the conclusions from Observation 2.8 and Theorem 2.10. That is, transfinite eliminations of non-best replies may be necessary to achieve consistency with common knowledge of rationality, but common knowledge of rationality is generally equal to the countable intersection of up to k-fold knowledge of rationality. However, since knowledge spaces have a different structure than the beliefs-type spaces I use here,[19] my characterization of common belief in rationality does not immediately translate to knowledge spaces merely by replacing belief with knowledge. A higher degree of comparability between knowledge- and belief-based results could be achieved within a

[19] In particular there exists no universal knowledge space, see Heifetz and Samet (1998a).

universal knowledge-belief space as in Meier (2008), and it seems plausible that extending my results to such a setting might lead to a simultaneous characterization of common belief in and common knowledge of rationality by means of transfinite elimination of non-best replies (Procedure 2.9) and ω-fold elimination of non-best replies and supporting beliefs (Procedure 3.2).

Another related paper is Bach and Cabessa (2012). They study common knowledge of rationality of rationality in infinite games, comparing it to so-called limit knowledge of rationality. Limit knowledge of rationality is defined as the limit of the sets of states exhibiting up to k-fold knowledge of rationality with respect to a pre-specified topology on the infinite state space. Bach and Cabessa's (2012) main result is that, for any subset of states exhibiting common knowledge of rationality, there is a topology on the state space such that limit knowledge under that topology selects the desired subset. Hence, limit knowledge gives a topological description of different ways in which one could refine common knowledge of rationality. Since Bach and Cabessa (2012) thus talk about connections between different solution concepts, their results do not concern the procedural characterization of common belief in rationality. In particular, instances where limit knowledge of rationality refines common knowledge of rationality are not logically connected to occurrences of transfinite elimination of non-best replies.

5 Conclusion

In this paper, I have provided a general belief-based analysis of transfinite elimination of non-best replies within the purely measurable beliefs-type space (Heifetz and Samet 1998b). As it turned out, transfinite elimination rounds are entirely explained by information loss that is incurred while characterizing the behaviors consistent with common belief in rationality by means of elimination of non-best replies. In particular, players in infinite games need not be endowed with transfinite reasoning depths to capture any reasoning relating to transfinite elimination of non-best replies.

With regards to the generality of these observations, note that two assumptions drive the results presented in this paper. For one, players reason about infinite sets of utility-relevant states using σ-additive probability measures. Separable spaces as in Definition 2.1 are the richest class of strategy sets that can be handled under this assumption.[20] For another, like in finite games, I consider players with naturalistic reasoning capabilities in the sense that all finite orders of beliefs $b_i^1, b_i^2, \ldots$ fully determine the belief hierarchy b_i of every player i.[21]

[20] Richer strategy sets could be accommodated if one assumed that beliefs are finitely additive measures as in Meier (2006). This would not lead to a one-size-fits-all solution for non-separable spaces, though: As Meier (2006) proves in his Theorem 4, there is no universal type space for the case of finitely-additive measures on fields that are closed under arbitrary intersections of events. In particular, there is no universal type space for the case of finitely-additive measures on the full power set of a given infinite state space. So my assumption of σ-additive beliefs does not cause the general constraint that *some* subsets of the event space cannot be reasoned about by players in some infinite games.

[21] Pinter (2019) considers σ-additive beliefs forming belief hierarchies of countably ordinal length. For this model, Observation 2.8 would break down, and, for any countable ordinal α, a strategy would survive $\alpha+1$-fold elimination of non-best replies iff it were consistent with α-fold belief in rationality. Furthermore, it seems natural to expect that some version of Theorem 2.9, Part 1 would imply that *uncountably* transfinite iterations of elimination of non-best

These two fundamental assumptions seem natural to impose on infinite games whenever one aims to gain general insights about large finite games (e.g. in the applications of infinite games in auction theory, bargaining theory, and industrial organization that were mentioned earlier). As such, my results can be regarded as a definite answer regarding how transfinite eliminations of non-best replies ought to be interpreted in any naturalistic model of strategic interaction that considers infinite games.

Appendix

A Proofs

Proof of Lemma 2.3

Proof. Let Γ be a static game and let $i \in I$. Then since S_i is separable, it contains a countably dense subset Q. For every $(s_i, b_i) \in S_i$ and $q \in Q$, define $D_q(s_i, b_i) := U_i(s_i, b_i) - U_i(q, b_i)$. Since $U_i : S_i \to \mathbb{R}$ is bounded and measurable, D_q is measurable.[22] Hence, for every $q \in Q$, the set $A_q = \{(s_i, b_i) | D_q(s_i, b_i) \geq 0\}$ is measurable. To finish the proof, it remains to show that $\bigcap_{q \in Q} A_q = R_i(1)$. The direction $\bigcap_{q \in Q} A_q \subseteq R_i(1)$ is clear. For the reverse direction, take any $(s_i, b_i) \in \bigcap_{q \in Q} A_q$, and note that $\inf_{q \in \omega} D_q(s_i, b_i) \geq 0$. But then, for any sequence $(q_k) \in Q^{\mathbb{N}}$, we also have $\underline{\lim}_{k \to \infty} D_{q_k}(s_i, b_i) \geq \inf_{q \in \omega} D_q(s_i, b_i) \geq 0$, and – using that Q is dense in S_i – it follows that $U_i(s_i, b_i) - U_i(s_i', b_i) \geq 0, \ \forall s_i' \in S_i$. Thus $(s_i, b_i) \in R_i(1)$, concluding the proof. □

Proof of Theorem 2.6

Proof. $\Rightarrow$: We start by showing that $s_i \in BR_i(k+1)$ if there exists $b_i \in B_i$ such that $(s_i, b_i) \in R_i(k+1)$. We proceed by induction over $k \geq 0$.

Induction Start: Suppose $(s_i, b_i) \in R_i(1)$ for some $b_i \in B_i$. Then the first-order belief $b_i^1 = \text{marg}_{S_{-i}} b_i$ must satisfy $U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \ \forall s_i' \in S_i$. It follows that $s_i \in BR_i(1)$.

Induction Step: Assume that, for all players i, $s_i \in BR_i(k+1)$ whenever there exists $b_i \in B_i$ such that $(s_i, b_i) \in R_i(k+1)$. Now let $(s_i, b_i) \in R_i(k+2)$. Then, since $R_i(k+2) \subseteq R_i(1)$, we have $U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \forall s_i' \in S_i$ where $b_i^1 = \text{marg}_{S_{-i}} b_i$. And furthermore, $b_i \in \Delta(R_{-i}(k+1))$ assigns full measure to the set of (s_{-i}, b_{-i})-tuples, where, for every $j \neq i$, $(s_j, b_j) \in R_j(k+1)$. Now, by the induction assumption, for every such (s_j, b_j), we have that $s_j \in BR_j(k+1)$, $j \neq i$. Hence, letting Σ denote the product σ-algebra over S_{-i}, there must exist $\mu \in \Delta(BR_{-i}(k+1))$ such that $b_i^1(E) = \mu(E \cap BR_{-i}(k+1))$ for every $E \in \Sigma$.

It follows that $s_i \in BR_i(k+2)$, establishing the first direction.

$\Leftarrow$: For this direction, we show that, for any $s_i \in BR_i(k+1)$, there is a belief hierarchy $b_i \in B_i$ such

replies have similar properties in this model as *all* transfinite iterations of elimination of non-best replies have under my assumptions.

[22] To show this formally, define $\phi_q : B_i \to S_i \times B_i$ via $\phi_q(b_i) = (q, b_i)$. Then ϕ_q is measurable, and (hence) $U_i \circ \phi_q$ is measurable. Now, fixing $q \in S_i$ and taking $U_i(q, \cdot)$ as a function on $S_i \times B_i$, for any Borel-measurable $R \subseteq \mathbb{R}$, we have $U_i^{-1}(q, R) = S_i \times \{b_i \in B_i | \ (U_i \circ \phi_q)(b_i) \subseteq R\}$, and it follows that $U_i(q, \cdot)$ is measurable.

that $(s_i, b_i) \in R_i(k+1)$. Again, we proceed by induction over $k \geq 0$.

Induction Start: Let $s_i \in BR_i(1)$. Then there is $b_i^1 \in B_i^1$ such that $U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1),\ \forall s_i' \in S_i$. So take any $b_i \in B_i$ such that $\text{marg}_{S_{-i}} b_i = b_i^1$. Then $(s_i, b_i) \in R_i(1)$.

Induction Step: Assume that, for every player i and any $s_i \in BR_i(k+1)$, there is a belief hierarchy $b_i \in B_i$ such that $(s_i, b_i) \in R_i(k+1)$. We have to show that if $s_i \in BR_i(k+2)$, then there is a belief hierarchy $b_i \in B_i$ such that $(s_i, b_i) \in R_i(k+2)$.

So let $s_i \in BR_i(k+2)$. Then, letting Σ denote the product σ-algebra on S_{-i}, there exists a measure $\mu \in \Delta(BR_{-i}(k+1))$ associated with a unique first-order belief $b_i^1 \in B_i^1$ such that $b_i^1(E) = \mu(E \cap BR_{-i}(k+1))$ for every $E \in \Sigma$ and $U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1),\ \forall s_i' \in S_i$. Furthermore, by the induction assumption, for any $j \neq i$ and $s_j \in BR_j(k+1)$, there is a belief hierarchy $\hat{b}_j(s_j) \in B_j$ such that that $(s_j, \hat{b}_j(s_j)) \in R_j(k+1)$. Let $\Theta = \{(s_{-i}, \hat{b}_{-i}(s_{-i})) \in S_{-i} \times B_{-i} |\ s_{-i} \in BR_{-i}(k+1)\}$, noting that $\Theta \subseteq R_{-i}(k+1)$ by construction. Now, letting $\mathcal{T}$ denote the product σ-algebra over $S_{-i} \times B_{-i}$, let $\hat{\mu} \in \Delta(\Theta, \mathcal{T}|\Theta)$ be the unique measure satisfying $\hat{\mu}(\hat{F}) = \mu(F)$ for $F = E \cap BR_{-i}(k+1) \in \Sigma | BR_{-i}(k+1)$ and every $\hat{F} = \hat{E} \cap \Theta \in \mathcal{T}|\Theta$ such that $\text{proj}_{S_{-i}} \hat{E} = E$. Let b_i be the unique belief hierarchy given by $b_i(\hat{E}) = \hat{\mu}(\hat{E} \cap \Theta)$ for every $\hat{E} \in \mathcal{T}$. By construction, $\text{marg}_{S_{-i}} b_i = b_i^1$ and $b_i \in \Delta(R_{-i}(k+1))$.

It follows that $(s_i, b_i) \in R_i(k+2)$, completing the second direction and hence the proof. □

Proof of Observation 2.8

Proof. $R_i(\omega+1) \subseteq R_i(\omega)$ is clear. To prove $R_i(\omega+1) \supseteq R_i(\omega)$ let $(s_i, b_i) \in R_i(\omega)$. By definition, s_i is rational for b_i and b_i assigns full measure to opponents' (s_{-i}, b_{-i})-tuples, expressing rationality and up to k-fold belief in rationality for all finite $k \geq 1$. Hence $b_i \in \Delta(R_{-i}(\omega))$, and it follows that $(s_i, b_i) \in R_i(\omega+1) = R_i(1) \cap (S_i \times \Delta(R_{-i}(\omega)))$. □

Proof of Theorem 2.10

Proof. **Part 1:** For any player $i \in I$, let $\overline{BR}_i \subseteq S_i$ denote the set of player i's strategies surviving transfinite elimination of non-best replies, recalling that this set exists as a consequence of the well-ordering theorem.

First, I show $\overline{BR}_i \supseteq \text{proj}_{S_i} R_i(\omega)$. Let $s_i \in \text{proj}_{S_i} R_i(\omega)$. Then there exists a belief hierarchy $b_i \in \Delta(R_{-i}(\omega))$ that rationalizes s_i. I now proceed by transfinite induction to prove that $s_i \in BR_i(\alpha)$ for every ordinal α. Using Theorem 2.6, it already follows that $s_i \in BR_i(k)$ for all finite $k \geq 1$ and (hence) that $s_i \in \bigcap_{k \in \omega} BR_i(k) = BR_i(\omega)$. Now for any ordinal α, assume that $s_i \in \text{proj}_{S_i} R_i(\omega)$ implies $s_i \in BR_i(\beta)$ for every $\beta < \alpha$. If α is a limit ordinal, then we immediately have $s_i \in BR_i(\alpha)$. Otherwise, if α is a successor ordinal, note that $b_i \in \Delta(R_{-i}(\omega))$ assigns full measure to tuples (s_{-i}, b_{-i}) such that $b_j \in \Delta(R_j(\omega))$ rationalizes s_j for every $j \neq i$. So letting Σ denote the product σ-algebra over S_{-i}, the induction assumption implies that $b_i^1 = \text{marg}_{S_{-i}} b_i$ must satisfy $b_i^1(E) = \mu(E \cap BR_{-i}(\alpha - 1))$ for some $\mu \in \Delta(BR_{-i}(\alpha - 1))$ and every $E \in \Sigma$. It follows that $s_i \in BR_i(\alpha)$, thus completing the transfinite induction.

Next, for $\overline{BR}_i \subseteq \text{proj}_{S_i} R_i(\omega)$, note that, by construction, we must have

$$\overline{BR}_i = \{s_i \in S_i |\ \exists b_i^1 \in B_i^1 \text{ s.th. } U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \forall s_i' \in S_i \text{ and } \exists \mu \in \Delta(\overline{BR}_{-i})$$

s.th. $b_i^1(E) = \mu(E \cap \overline{BR}_{-i})$ for every measurable $E \subseteq S_{-i}\}$.

Otherwise, further eliminations of non-best replies would be possible.

In the following, let Σ_{-i}^0 denote the product σ-algebra over S_{-i} and, for every $k \geq 1$, let Σ^k denote the product σ-algebra over $S_{-i} \times \Delta(B_{-i}^k)$. Also define $\overline{BR}_{-i}^0 := \overline{BR}_{-i}$. Take $s_i \in \overline{BR}_i$ and recursively construct a sequence of finite-order beliefs $b_i^1,\ b_i^2, \cdots \in \times_{k \in \omega} B_i^k$ as follows:

Step 1: Let $b_i^1 \in B_i^1$ be such that $U_i(s_i, b_i^1) \geq U_i(s_i', b_i^1), \forall s_i' \in S_i$ and $b_i^1(E) = \mu^1(E \cap \overline{BR}_{-i}^0)$ for some $\mu^1 \in \Delta(\overline{BR}_{-i}^0)$ and every $E \in \Sigma_{-i}^0$. Analogously, for every $j \in I$ and every $s_j \in \overline{BR}_j^0$ fix some $b_j^1[s_j] \in B_j^1$ such that $U_j(s_j, b_j^1[s_j]) \geq U_j(s_j', b_j^1[s_j]), \forall s_j' \in S_j$ and $b_j^1[s_j](E) = \mu_j^1[s_j](E \cap \overline{BR}_{-j}^0)$ for some $\mu_j^1[s_j] \in \Delta(\overline{BR}_{-j}^0)$ and all $E \in \Sigma_{-j}^0$. Define $\overline{BR}_j^1 := \{(s_j, b_j^1[s_j]) |\ s_j \in \overline{BR}_j^0\}$.

Step $k > 1$: Assume b_i^{k-1} and $\overline{BR}_j^{k-1}, j \in I$ are defined. Take $\mu_i^k \in \Delta(\overline{BR}_{-i}^{k-1})$ such that $\text{marg}_{\overline{BR}_{-i}^{k-2}} \mu_i^k = \mu_i^{k-1}$ and let $b_i^k \in B_i^k$ be the unique measure such that $b_i^k(E) = \mu_i^k(E \cap \overline{BR}_{-i}^{k-1})$ for all $E \in \Sigma_{-j}^{k-1}$.[23] Analogously, for every $j \in I$ and every $(s_j, b_j^{k-1}[s_j]) \in \overline{BR}_j^{k-1}$, take $\mu_j^k[s_j] \in \Delta(\overline{BR}_{-j}^{k-1})$ such that $\text{marg}_{\overline{BR}_{-j}^{k-2}} \mu_j^k[s_j] = \mu_j^{k-1}[s_j]$ and let $b_j^k[s_j] \in \Delta(S_{-j} \times B_{-j}^{k-1})$ be the unique measure such that $b_j^k[s_j](E) = \mu_j^k(E \cap \overline{BR}_{-j}^{k-1})$ for all $E \in \Sigma_{-j}^{k-1}$. To complete the construction, define $\overline{BR}_j^k := \{(s_j, b_j^k[s_j]) |\ s_j \in \overline{BR}_j^0\}$.

By construction, $b_i^1, b_2, \ldots$ is a coherent sequence of finite-order beliefs in $\times_{k \in \omega} B_i^k$. Using Proposition 5.4 in Heifetz and Samet (1999), it then follows that there exists a belief hierarchy $b_i \in B_i$ such that $b_i^k = \text{marg}_{S_{-i} \times B_{-i}^{k-1}} b_i$ for every $k \geq 1$. For further reference, note that also the sequences $b_j^1[s_j], b_j^2[s_j], \ldots$ we constructed for every $j \in I$ and $s_j \in \overline{BR}_j$ must each induce a belief hierarchy $b_j[s_j] \in B_j$ by implication.

I will now prove, by induction, that $(s_i, b_i) \in R_i(\omega)$. $(s_i, b_i) \in R_i(1)$ immediately follows from Step 1 of the construction above, and the same goes for $(s_j, b_j[s_j]) \in R_j(1)$ for every $s_j \in \overline{BR}_j$. Now for any $k \geq 1$, assume that $(s_i, b_i) \in R_i(k)$ and $(s_j, b_j[s_j]) \in R_j(k)$ for every $s_j \in \overline{BR}_j$. Then, by construction, $b_i \in \Delta(R_{-i}(k))$, and it follows that $(s_i, b_i) \in R_i(k+1)$, completing the induction, and hence the proof of Part 1.

Part 2: $BR_i(\omega) \supseteq \overline{BR}_i$ is clear. To show that the reverse direction also applies under the compact-continuity assumptions of Part 2, assume that $s_i \in BR_i(\omega)$. We will show that $s_i \in \text{proj}_{S_i} R_i(\omega)$. Since $s_i \in BR_i(\omega)$, it must be consistent with up to k-fold belief in rationality for any fixed $k \geq 0$ by Theorem 2.6. For any $k \geq 0$, let $B_i[k, s_i]$ be the set of belief hierarchies that rationalize s_i and express up to k-fold belief in rationality. I will now show that $B_i[k, s_i]$ is a compact set for every

[23] That a unique measure b_i^k with those properties exists and that it lies in B_i^k can be seen as follows: With the induction start, for every $j \in I$ and every $s_j \in \overline{BR}_j$, we have $b_j^1[s_j] \in B_j^1$. So let $k > 1$ assume that $b_j^{k-1}[s_j] \in B_j^{k-1}$ for every $j \in I$ and every $s_j \in \overline{BR}_j$. Noting that $\overline{BR}_{-i}^{k-1} \subseteq S_{-i} \times B_{-i}^{k-1}$ for every $k \geq 1$, the existence of a measure $b_i^k \in \Delta(S_{-i} \times B_{-i}^{k-1})$ then again follows from Lemma 2.2 in Heifetz and Samet (1999). Finally, to see that $b_i^k \in B_i^k$, note that, for every for every $j \in I$ and every $s_j \in \overline{BR}_j$, $b_j^{k-1}[s_j] \in B_j^{k-1}$ induces some belief hierarchy $b_j[s_j] \in B_j$. Define $\Theta := \{(s_{-i}, b_{-i}[s_{-i}]) \in S_{-i} \times B_{-i} |\ s_{-i} \in \overline{BR}_{-i}\}$. Now letting $\mathcal{T}$ denote the product σ-algebra over $S_{-i} \times B_{-i}$, let $\hat{\mu} \in \Delta(\Theta, \mathcal{T}|\Theta)$ be the unique measure satisfying $\hat{\mu}(\hat{F}) = \mu^k(F)$ for $F = E \cap \overline{BR}_{-i}^{k-1}$ and $\hat{F} = \hat{E} \cap \Theta$ such that $\text{proj}_{S_{-i} \times B_{-i}^{k-1}} \hat{F} = F$. Now let $b_i \in B_i = \Delta(S_{-i} \times B_{-i})$ be the unique measure such that $b_i(\hat{E}) = \hat{\mu}(\hat{E} \cap \Theta)$ for every $\hat{E} \in \mathcal{T}$. Since $b_i^k = \text{proj}_{S_{-i} \times B_{-i}^{k-1}} b_i$, it follows that $b_i^k \in B_i^k$, completing the induction.

$k \geq 0$. Since the sequence $B_i[0, s_i], B_i[1, s_i], \ldots$ is then a decreasing sequence of nested non-empty compact sets, Cantor's intersection theorem implies that $B_i[\omega, s_i] = \bigcap_{k \in \{0,1,2,\ldots\}} B_i[k, s_i]$ is non-empty such that indeed $s_i \in \text{proj}_{S_i} R_i(\omega)$. It remains to prove, by induction over $k \geq 0$, that every $B_j[k, s_j]$ is compact and metrizable for every player j, every $s_j \in S_j$ and every $k \geq 0$:

Induction Start: Take $b_j \in B_j \backslash B_j[0, s_j]$. Then s_j is not rational given b_j. Hence, by belief continuity, there is an open set $\hat{B}_j \subseteq B_j \backslash B_j[0, s_j]$ such that s_j is not rational given any $\hat{b}_j \in \hat{B}_j$. It follows that $B_j \backslash B_j[0, s_j]$ is open and, consequently, $B_j[0, s_j]$ is closed. Since B_j is compact and metrizable,[24] $B_j[0, s_j]$ is compact and metrizable.

Induction Step: Assume that $B_j[k, s_j]$ is compact and metrizable for any player j, any $s_j \in S_j$, and for some $k \geq 0$. We can write

$$B_j[k+1, s_j] = B_j[k, s_j] \cap \Delta\left(\bigtimes_{\ell \neq j} \left\{(s_\ell, b_\ell) \middle| \, b_\ell \in B_\ell[k, s_\ell]\right\}\right)$$

By the induction assumption, $\times_{\ell \neq j}\{(s_\ell, b_\ell)|\, b_\ell \in B_\ell[k, s_\ell]\}$ is compact and metrizable. Since the set of probability measures over a compact and metrizable set is itself compact and metrizable, the same is true for $\Delta(\times_{\ell \neq j}\{(s_\ell, b_\ell)|\, b_\ell \in B_\ell[k, s_\ell]\})$. It follows that $B_j[k+1, s_j]$ is compact and metrizable, completing the induction and hence the proof. □

Proof of Theorem 3.3

Proof. To prove (1), we show that $\overline{R}_i^*(k) = R_i(k)$ for all $k \in \omega$ and all players i. Recall that $R_i(k)$ is the set of (c_i, b_i) tuples expressing rationality and up to $k-1$-fold belief in rationality for $k \geq 1$ (Definition 2.4). (2) then directly follows from the definition of common belief in rationality (Definition 2.7). We prove $\overline{R}_i^*(k) = R_i(k)$ by induction over $k \geq 1$.

Induction Start: For $k = 1$, the statement follows directly from the fact that utility depends on choices and 1st-order beliefs.

Induction Step: Let $R_i(k) = \overline{R}_i^*(k)$ for $k \geq 1$ and all players i. Then

$$\begin{aligned}
R_i(k+1) =& R_i(1) \cap (S_i \times \Delta(R_{-i}(k))) \\
=& \{(s_i, b_i) \in R_i(k) |\, b_i \in \Delta(R_{-i}(k))\} \\
=& \{(s_i, b_i) \in \overline{R}_i^*(k) |\, b_i \in \Delta(\overline{R}_{-i}^*(k))\} \\
=& \{(s_i, b_i) \in S_i \times B_i |\, (s_i, b_i^k) \in R_i^*(k) \text{ and } b_i \in \Delta(\overline{R}_{-i}^*(k))\} \\
=& \left\{(s_i, b_i) \in S_i \times B_i |\, (s_i, b_i^k) \in R_i^*(k) \text{ and } b_i \in \Delta\left(\left\{(s_{-i}, b_{-i}) \in S_{-i} \times B_{-i} | (s_{-i}, b_{-i}^k) \in R_{-i}^*(k)\right\}\right)\right\} \\
=& \{(s_i, b_i) \in S_i \times B_i |\, (s_i, b_i^k) \in R_i^*(k) \text{ and } b_i^{k+1} \in \Delta(R_{-i}^*(k))\} \\
=& \{(s_i, b_i) \in S_i \times B_i |\, (s_i, b_i^{k+1}) \in R_i^*(k+1)\} \\
=& \overline{R}_i^*(k+1).
\end{aligned}$$

The induction, and hence the proof, is now complete. □

[24] For each player i, compactness and metrizability of B_i follows from the compactness and metrizability of all S_j, $j \in I$ and Tychonoff's theorem.

B A Belief-Based Analysis for Infinite Psychological Games

In static psychological games, utility is allowed to depend on arbitrary levels of higher-order beliefs $b_i^n \in B_i^n$, $n \geq 1$ or even on the full belief hierarchy $b_i \in B_i$.

Definition B.1. *(Static Psychological Game)*
A static psychological game is a tuple $\Gamma = (S_i, B_i, U_i)_{i \in I}$ *with* I *an arbitrary set of players,* S_i *a separable set of strategies for player* i*,* B_i *the set of belief hierarchies for player* i *expressing coherency and common belief in coherency,*[5] *and* U_i *a measurable and bounded utility function* $U_i : S_i \times B_i \to \mathbb{R}$*.*

In static psychological games, dependence of utility on higher-order beliefs makes it impossible to characterize rational choice and hence common belief in rationality only based on strategy elimination. However, there is an intuitive generalization of the fundamental theorem of epistemic game theory: Take a psychological game Γ that is *belief-finite of order* $n+1$ for some $n \geq 1$. I.e. $U_i : S_i \times B_i^{n+1} \to \mathbb{R}$ for all players $i \in I$. Then a tuple $(s_i, b_i^n) \in S_i \times B_i^n$ is consistent with common belief in rationality iff it survives *transfinite elimination of strategies and nth-order beliefs.* Jagau and Perea (2017) prove this characterization for the special case of static psychological games with finite sets of strategies.

Here, I extend Theorem 2.10 above to yield a general version of Jagau and Perea's (2017) result for infinite psychological games as in Definition B.

In preparation for the result, first note that Definitions 2.2, 2.4, and 2.7 for belief in rationality immediately translate to static psychological games. The only change we need to make is to replace the traditional utility function $U_i : S_i \times B_i^1 \to \mathbb{R}$ by a psychological utility function $U_i : S_i \times B_i \to \mathbb{R}$. As a consequence, also Lemma 2.3 extends after substituting psychological utility functions and Observation 2.8 extends verbatim. This establishes that all layers of rationality and belief in rationality are measurable and that common belief in rationality is the terminal depth of reasoning – also in psychological games.

Next, we can define *transfinite elimination of strategies and nth-order beliefs*:

Procedure B.2. *(Transfinite Elimination of Strategies and nth-Order Beliefs)*[25]
Step 1*: For every player* $i \in I$*, let*

$$BR_i^n(1) = \{(s_i, b_i^n) \in S_i \times B_i^n \mid \exists b_i^{n+1} \in B_i^{n+1} \text{ s.th. } U_i(s_i, b_i^{n+1}) \geq U_i(s_i', b_i^{n+1}), \forall s_i' \in S_i$$
$$\text{and } \text{marg}_{S_{-i} \times B_{-i}^{n-1}} b_i^{n+1} = b_i^n\}.$$

Step $k > 1$*: Assume* $BR_i^n(k-1)$ *is defined for every player* i*. Then, for every player* i*,*

$$BR_i^n(k) = \{(s_i, b_i^n) \in S_i \times B_i^n \mid \exists b_i^{n+1} \in B_i^{n+1} \text{ s.th. } U_i(s_i, b_i^{n+1}) \geq U_i(s_i', b_i^{n+1}), \forall s_i' \in S_i,$$
$$\text{s.th. } \text{marg}_{S_{-i} \times B_{-i}^{n-1}} b_i^{n+1} = b_i^n, \text{ and } \exists \mu \in \Delta(BR_{-i}^n(k-1)) \text{ s.th.}$$
$$b_i^{n+1}(E) = \mu(E \cap BR_{-i}^n(k-1)) \text{ for every measurable } E \subseteq S_{-i} \times B_{-i}^n\}.$$

[25] Note that setting $n = 0$ yields elimination of non-best replies as defined in the main text.

Step $\alpha \geq \omega$: Assume $BR_i^n(k)$ is defined for every player i and every $k \in \omega$.
Now, for every player i, let $BR_i^n(\omega) = \bigcap_{k\in\omega} BR_i^n(k)$.
Furthermore, for every successor ordinal $\alpha > \omega$, let

$$BR_i^n(\alpha) = \{(s_i, b_i^n) \in S_i \times B_i^n \mid \exists b_i^{n+1} \in B_i^{n+1} \text{ s.th. } U_i(s_i, b_i^{n+1}) \geq U_i(s_i', b_i^{n+1}), \forall s_i' \in S_i,$$
$$\text{s.th. } \operatorname{marg}_{S_{-i}\times B_{-i}^{n-1}} b_i^{n+1} = b_i^n, \text{ and } \exists \mu \in \Delta(BR_{-i}^n(\alpha-1))$$
$$\text{s.th. } b_i^{n+1}(E) = \mu(E \cap BR_{-i}^n(\alpha-1)) \text{ for every measurable } E \subseteq S_{-i} \times B_{-i}^n\}.$$

Lastly, for every limit ordinal $\alpha \geq \omega$, define $BR_i^n(\alpha) = \bigcap_{\beta<\alpha} BR_i^n(\beta)$.

Inspecting Procedure B.2, it is interesting to note that we here encounter an analogous non-measurability issue for (s_i, b_i^n)-tuples as the one we encountered with respect to strategies in Definition 2.5. So rather intuitively, the domain of utility functions u_i generally constrains which strategy-belief combinations can assume a meaningful role in players' strategic reasoning. In particular, for any belief-finite games of order $n+1$, only strategy-kth-order-belief tuples of order $k \geq n+1$ will meaningfully figure in players' theories about the game.

With these preliminaries, we get a straightforward generalization of Theorems 2.6 and 2.10.

Theorem B.3. *(Consistency with up to k-Fold and Common Belief in Rationality)*
Let Γ be a static psychological game that is belief-finite of order $n+1$ for some $n \geq 0$.

1. *For any player i, any $(s_i, b_i^n) \in S_i \times B_i^n$, and any finite $k \geq 0$, there exists a belief hierarchy $b_i \in B_i$ such that $\operatorname{marg}_{S_{-i}\times B_{-i}^{n-1}} b_i = b_i^n$ and $(s_i, b_i) \in R_i(k+1)$ iff $s_i \in BR_i^n(k+1)$.*

2. *For any player i and any $(s_i, b_i^n) \in S_i \times B_i^n$, there exists $b_i \in B_i$ such that $\operatorname{marg}_{S_{-i}\times B_{-i}^{n-1}} b_i = b_i^n$ and $(s_i, b_i) \in R_i(\omega)$ iff (s_i, b_i^n) survives transfinite elimination of non-best replies.*

3. *For all players i, let S_i be compact Haussdorf. Moreover, let U_i be such that, for every player i, every $s_i \in S_i$, every $n+1$th-order belief b_i^{n+1}, and every $\varepsilon > 0$, there is a $\delta > 0$ such that for any belief $\hat{b}_i^{n+1}$ with $d(b_i^{n+1}, \hat{b}_i^{n+1}) < \delta$ we have $|U_i(s_i, b_i^{n+1}) - U_i(s_i, \hat{b}_i^{n+1})| < \varepsilon$. Then (s_i, b_i^n) survives transfinite elimination of non-best replies iff $(s_i, b_i^n) \in BR_i^n(\omega)$.*

The proof of Theorem B.3 is entirely analogous to the proofs of Theorems 2.6 and 2.10. One only needs to replace strategies s_i with (s_i, b_i^n)-tuples and substitute suitable σ-algebras and $BR_{-i}^n(k)$- and $\overline{BR}_{-i}^n$-restrictions thereof accordingly.

Lastly, note that also the exact characterization of common belief in rationality has an appealing generalization to psychological games. For any $n \geq 1$ and any psychological game that is belief finite of order $n+1$, the only necessary change to Procedure 3.2 is to select (s_i, b_i^{n+1})-tuples at Step 1, to select (s_i, b_i^{n+2})-tuples at step 2, and so on. Making analogous substitutions in the proof of Theorem 3.3 then shows that the so-constructed *generalized elimination of non-best replies and supporting beliefs* again gives an exact characterization of common belief in rationality. Note that, intuitively, letting $n \to \infty$, the so-defined procedure reverts back to Definitions 2.4 and 2.7 for up to k-fold and common belief in rationality.

Bibliography

Apt, K. R., 2007: The many faces of rationalizability. *The B.E. Journal of Theoretical Economics*, **7 (1)**.

Arieli, I., 2010: Rationalizability in continuous games. *Journal of Mathematical Economics*, **46**, 912–924.

Aumann, R. J., 1976: Agreeing to disagree. *The Annals of Statistics*, **4 (6)**, 1236–1239.

Bach, C., and J. Cabessa, 2012: Common knowledge and limit knowledge. *Theory and Decision*, **73 (3)**, 423–440.

Bernheim, B. D., 1984: Rationalizable strategic behavior. *Econometrica*, **52 (4)**.

Brandenburger, A., 2014: *The Language of Game Theory: Putting Epistemics into the Mathematics of Games*, Vol. 5. World Scientific Series in Economic Theory.

Brandenburger, A., and E. Dekel, 1987: Rationalizability and correlated equilibria. *Econometrica*, **55 (6)**, 1391–1402.

Brandenburger, A., and E. Dekel, 1993: Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*, **59 (1)**, 189–198.

Burali-Forti, C., 1897: Una questione sui numeri transfiniti. *Rendiconti del Circolo Matematico di Palermo*, **11**, 154–164.

Chen, Y.-C., N. V. Long, and X. Luo, 2007: Iterated strict dominance in general games. *Games and Economic Behavior*, **61 (2)**, 299–315.

Dufwenberg, M., and M. Stegeman, 2002: Existence and uniqueness of maximal reductions under iterated strict dominance. *Econometrica*, **70 (5)**, 2007–2023.

Heifetz, A., and D. Samet, 1998a: Knowledge spaces with arbitrarily high rank. *Games and Economic Behavior*, **22 (2)**, 260–273.

Heifetz, A., and D. Samet, 1998b: Topology-free typology of beliefs. *Journal of Economic Theory*, **82 (2)**, 324–341.

Heifetz, A., and D. Samet, 1999: Coherent beliefs are not always types. *Journal of Mathematical Economics*, **32**, 475–488.

Jagau, S., and A. Perea, 2017: Expectation-based psychological games and psychological expected utility, working paper.

Lipman, B. L., 1994: A note on the implications of common knowledge of rationality. *Games and Economic Behavior*, **6 (1)**, 114–129.

Meier, M., 2006: Finitely additive beliefs and universal type spaces. *The Annals of Probability*, **34 (1)**, 386–422.

Meier, M., 2008: Universal knowledge–belief structures. *Games and Economic Behavior*, **62 (1)**, 53–66.

Milgrom, P., and J. Roberts, 1990: Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, **58 (6)**.

Pearce, D., 1984: Rationalizable strategic behavior and the problem of perfection. *Econometrica*, **52 (4)**, 1029–1050.

Pinter, M., 2010: The non-existence of a universal topological type space. *Journal of Mathematical Economics*, **46**, 223–229.

Pinter, M., 2019: A new epistemic model, working paper available at https://drive.google.com/file/d/1Ys6Ozs2mosXrK1FOO4yHnqeuUHdlCyAB/view.

Samet, D., 2015: Is common knowledge of rationality sluggish?, working paper available at https://www.tau.ac.il/~samet/papers/sluggish-ck.pdf.

Tan, T. C.-C., and S. R. da Costa Werlang, 1988: The bayesian foundations of solution concepts of games. *Journal of Economic Theory*, **45 (2)**, 370–391.