Journal of Mathematical Economics 100 (2022) 102635

Contents lists available at ScienceDirect

Journal of Mathematical Economics

journal homepage: www.elsevier.com/locate/jmateco

Common belief in rationality in psychological games
Belief-dependent utility and the limits of strategic reasoning

Stephan Jagau [a,b,*], Andrés Perea [b]

[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

ARTICLE INFO

Article history:
Received 16 February 2021
Received in revised form 4 October 2021
Accepted 3 January 2022
Available online 22 January 2022
Manuscript handled by Editor Qingmin Liu

Keywords:
Psychological games
Belief-dependent motivation
Strategic rationality
Common belief in rationality
Rationalizability
Epistemic game theory

ABSTRACT

Psychological games enable us to study diverse motivations like anger, guilt, and intention-based reciprocity using models of rational strategic choice based on common belief in rationality (aka correlated rationalizability). This is achieved by letting utility depend not only on outcomes and beliefs about others' behavior but also on higher-order beliefs. It is an open question whether such belief-dependent utilities can be made consistent with common belief in rationality in all empirically relevant cases. In this paper, we use a novel existence condition to show that common belief in rationality is possible for any empirically relevant case of belief-dependent utility. In addition, we present a recursive elimination procedure that characterizes common belief in rationality under minimal assumptions on belief-dependent utility functions.

1. Introduction

Traditional game theory assumes that decision-makers exclusively care about the outcomes that materialize as a result of their choices and opponents' choices. However, in many real-life interactions, we see ourselves caring not only about outcomes, but also about beliefs, expectations, and emotional reactions. Outcome-based utility as used in traditional game theory gives us a hard time trying to capture this aspect of human behavior. *Psychological game theory* (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009) addresses this issue by allowing players' utilities to directly depend not only on their choices and beliefs about others' choices, but also on arbitrary levels of higher-order beliefs. Since their introduction, psychological games have found many applications in behavioral and experimental economics. They have been used to study a diverse set of belief-dependent motivations such as intention-based reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Sebald, 2010), guilt and surprise (Dufwenberg, 2002; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007; Khalmetski et al., 2015; Attanasi et al., 2016, 2019), social norms and conformity (Huck and Kübler, 2000; Li, 2008), anxiety (Caplin and Leahy, 2004), lying behavior (Dufwenberg and Dufwenberg, 2018), and anger (Battigalli et al., 2019).

A key distinction of psychological game theory relative to other accounts of emotions and related phenomena is that it subjects belief-dependent motivations to the logic of rational strategic choice. Just as in traditional game theory, we can analyze strategic reasoning under belief-dependent utilities using standard game-theoretic solution concepts based on common belief in rationality (Spohn, 1982; Brandenburger and Dekel, 1987; Tan and da Costa Werlang, 1988, characterizing correlated rationalizability). Battigalli and Dufwenberg (2009) were the first to present such an epistemic framework for strategic reasoning expressing common strong belief in rationality (Battigalli and Siniscalchi, 2002, characterizing extensive-form rationalizability, Pearce, 1984) in dynamic psychological games where utility may depend on the full system of conditional belief hierarchies.

Despite the broad scope of psychological games in existing theory and applications, an open foundational question remains: What assumptions about belief-dependent utility are necessary to apply models of strategic rationality in the first place? Most existing investigations of psychological games restrict to games with finite choice- and player-sets, and they assume continuity of utility functions in higher-order beliefs to ensure that common belief in rationality (or a suitable extension) is possible. Psychological games that lie beyond these assumptions have not been

* Corresponding author at: IMBS, University of California, Irvine, Social Science Plaza A, Irvine, CA 92697-5100, USA.
E-mail addresses: sjagau@uci.edu (S. Jagau), a.perea@maastrichtuniversity.nl (A. Perea).
URLs: https://sites.google.com/view/stephanjagau (S. Jagau), http://www.epicenter.name/Perea/ (A. Perea).

https://doi.org/10.1016/j.jmateco.2022.102635

as thoroughly investigated. In particular, it is unknown whether there are empirically relevant cases of psychological games where belief-dependent utility causes models of strategic rationality and related solution concepts to fail.

In this paper, we get to the root of this question by providing an in-depth analysis of how belief-dependent utility affects common belief in rationality — the basic building block in all models of strategic reasoning in games. For maximum clarity, we do not attempt to explicitly cover all extensions of psychological games that have been studied in previous theory and applications. Instead, we go back to the simplest setup in which the interaction of belief-dependent motivation and strategic reasoning can be studied: static psychological games as defined in Geanakoplos et al. (1989).[1] Within this restricted class of games, we relax the aforementioned conventional assumptions on psychological utility functions and choice- and player-sets as much as possible.[2]

Our investigation consists of two parts, each of which zooms in on a basic issue where strategic reasoning in psychological games transcends strategic reasoning in traditional games.

The first of these issues is the possibility of common belief in rationality. We present an elementary example of non-existence, accompanied by a novel sufficient condition for possibility of common belief in rationality in psychological games. Our condition, *preservation of rationality at infinity*, is considerably weaker than the previously known conditions (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009; Bjorndahl et al., 2013).[3] In light of the new sufficient condition, we can see that our example encapsulates the typical configuration where belief-dependent utility causes inconsistencies with common belief in rationality. Not only that, it is now easy to understand that impossibility of common belief in rationality cannot occur in empirically plausible applications of psychological games.

The second issue we consider is the procedural characterization of common belief in rationality. We present an elimination procedure for psychological games that generalizes iterated elimination of strictly dominated choices from traditional game theory in intuitive ways: If player's utilities depend on $n+1$th-order beliefs, we find that iterative elimination of choices and nth-order beliefs characterizes common belief in rationality. This nests the case of traditional static games where players' utilities depend on first-order beliefs and iterated elimination of choices characterizes common belief in rationality. Extending upon previously results (Battigalli et al., 2020),[3] we use transfinite elimination of choices and nth-order beliefs[4] to achieve a characterization for static psychological games under minimal assumptions on utility functions. In particular, we do not require any form of continuity.

Our characterization theorem is accompanied by a simple example that gives a clear intuition for why characterizing common belief in rationality requires elimination of choices *and* beliefs. In particular, we illustrate how the procedural characterization is intimately related to *interactions* between players' preferences over choices and rationality constraints of their opponents that can arise in psychological games but not in traditional ones.

While *iterated elimination of strictly dominated choices* for traditional games is both implementable as a linear program and converges in finitely many steps, neither of these nice properties is inherited by our procedure *iterated elimination of choices and nth-order beliefs*. Following our main investigation, we survey and discuss the principle ways in which elimination of choices and nth-order beliefs differs from elimination of strictly dominated strategies.

Notably, it is straightforward to extend the results presented in this paper to psychological games with infinite choice and/or player sets — a class of games that has rarely, if ever, been considered in previous literature. Following our main analysis, we discuss these extensions as well as how to extend our results to dynamic games and games under asymmetric information.

The remainder of this paper proceeds as follows: Section 2 introduces static psychological games. Section 3 defines common belief in rationality. Section 4 provides sufficient conditions for common belief in rationality to be possible in a given psychological game. Section 5 presents the procedure *iterated elimination of choices and nth-order beliefs*. Section 6 discusses extensions to our results. Section 7 discusses related literature. Section 8 concludes. All proofs are in Appendix A at the end of the paper.

2. Psychological games

In traditional games, utilities u_i of players $i \in I$ depend on their choices C_i and on first-order beliefs about opponents' choices $B_i^1 = \Delta(C_{-i})$. Moreover, utilities depend *linearly* on first-order beliefs. By contrast, utilities in psychological games might depend *non-linearly* on the full *belief hierarchy* of players. Each belief hierarchy b_i is a sequence of probability distributions $(b_i^1, b_i^2, \dots)$ that capture i's belief about his opponents' choices (b_i^1), i's beliefs about opponents' choices *and* their beliefs about their opponents' choices (b_i^2), and so on.

Throughout the analysis, we only consider belief hierarchies satisfying coherency and common belief in coherency.[5] Brandenburger and Dekel (1993) prove that every such belief hierarchy $b_i \in B_i$ is homeomorphic to a probability distribution over opponents' choices and belief hierarchies $\delta(b_i) \in \Delta(C_{-i} \times B_{-i})$.[6] As a corollary, looking at nth-order beliefs $B_i^n = \text{proj}_{C_{-i} \times B_{-i}^{n-1}} B_i$, we find that every $b_i^n \in B_i^n$ is homeomorphic to a probability distribution over opponents' choices and $n-1$th-order beliefs $\tilde{\delta}(b_i^n) \in \Delta(C_{-i} \times B_{-i}^{n-1})$. We identify b_i with $\delta(b_i)$ and b_i^n with $\tilde{\delta}(b_i^n)$ whenever that is useful.

Drawing on these preliminaries, we now give a formal definition of static psychological games[7]:

[1] Dynamic psychological games differ from static psychological games as considered here in two ways: Firstly, players are allowed to move sequentially. Secondly, preferences may depend on updated beliefs that arise during the play of the dynamic game. In Section 6.3, we discuss how our results extend to dynamic psychological games.

[2] Specifically, our most permissive set of assumptions (see Section 6.2) will consist of minimal assumptions that guarantee the existence of a universal type space (Heifetz and Samet, 1998, 1999) and the measurability of rationality as an event within this universal type space.

[3] See Section 7 for a literature review.

[4] The need for transfinite eliminations reveals an interesting connection to correlated rationalizability in infinite traditional games, see Comment 3 on Theorem 5.5 and Section 6.2 for details.

[5] Our construction of these belief hierarchies essentially follows the standard procedure from Brandenburger and Dekel (1993). Coherency requires that every belief hierarchy $(b_i^1, b_i^2, \dots)$ 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. That is, 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. Battigalli et al. (2020) proceed in a slightly different way in their construction, see Footnote 38.

[6] Here and in the following, for any measurable set S, $\Delta(S)$ will denote the set of σ-additive probability measures on the Borel σ-algebra over S.

[7] Static versus dynamic psychological games are not clearly delineated in the literature. In line with Geanakoplos et al. (1989), we choose to call a psychological game static iff

1. the game involves a single stage of simultaneous moves by all players,
2. players' utilities *exclusively* depend on their choices and on first- and higher-order beliefs regarding players' choices *at the time of choosing*.

Definition 2.1 (*Static Psychological Game*)**.** A static psychological game is a tuple $\Gamma = (C_i, B_i, u_i)_{i \in I}$ with I a finite set of players, C_i the finite set of choices available to player i, B_i the set of belief hierarchies for player i expressing coherency and common belief in coherency, and u_i a measurable and bounded utility function of the form

$$u_i : C_i \times B_i \to \mathbb{R}.$$

Our Definition 2.1 for static psychological games is slightly different from previous ones. In our online appendix, we compare it to the best-known alternative from Battigalli and Dufwenberg (2009). As we demonstrate, our definition is entirely equivalent to theirs.

3. Rationality and common belief in rationality

In this section we extend the traditional definition of common belief in rationality to static psychological games. As in the traditional case, we start with defining rational choice:

Definition 3.1 (*Rational Choice*)**.** Choice $c_i \in C_i$ is rational for player i given belief hierarchy $b_i \in B_i$ if $u_i(c_i, b_i) \geq u_i(c_i', b_i), \ \forall c_i' \in C_i$.

Building on Definition 3.1, we define up to k-fold belief rationality for any finite k and common belief in rationality: Belief hierarchies express 1-fold belief in rationality if they assign full probability to opponents' choice-belief combinations such that the choice is rational for the opponent's belief. They express up to 2-fold belief in rationality if they express 1-fold belief in rationality and assign full probability to opponents' beliefs expressing 1-fold belief in rationality. And so on.

Definition 3.2 (*Up to k-Fold and Common Belief in Rationality*)**.** Recursively define

$$R_i(1) = \{(c_i, b_i) \in C_i \times B_i |\ u_i(c_i, b_i) \geq u_i(c_i', b_i), \forall c_i' \in C_i\}$$

$$R_i(2) = \{(c_i, b_i) \in R_i(1) |\ b_i \in \Delta(\bigtimes_{j \neq i} R_j(1))\}$$

$$R_i(k) = \{(c_i, b_i) \in R_i(k-1) |\ b_i \in \Delta(\bigtimes_{j \neq i} R_j(k-1))\}, \ k > 1$$

$$R_i(\omega) = \bigcap_{k \in \{1,2,\dots\}} R_i(k).$$

A belief hierarchy b_i expresses up to k-fold belief in rationality if $b_i \in \Delta(\bigtimes_{j \neq i} R_j(k))$. It expresses common belief in rationality if $b_i \in \Delta(\bigtimes_{j \neq i} R_j(\omega))$.

A choice c_i is *rational under up to k-fold belief in rationality* if there exists a belief hierarchy b_i satisfying $(c_i, b_i) \in R_i(k+1)$. A choice-nth-order-belief tuple (c_i, b_i^n) is *consistent with up to k-fold belief in rationality* if there exists a belief hierarchy b_i inducing b_i^n and satisfying $(c_i, b_i) \in R_i(k+1)$. Rationality under common belief in rationality and consistency with common belief in rationality are defined analogously.

Definition 3.2 proceeds in a very similar fashion as in traditional games. The main difference is the more general definition of rationality and the resulting need to track belief hierarchies as we iteratively refine belief in rationality.

Like in traditional games, two questions around common belief in rationality arise. The first one is whether for every psychological game Γ and every player i in it, there is a belief hierarchy b_i that expresses common belief in rationality. The second one is whether there is a procedure that identifies *all* choices that are rational under common belief in rationality for a player i. We investigate the first question in Section 4 and the second question in Section 5.

4. When common belief in rationality is possible

In this section we explore the conditions under which *common belief in rationality* is possible in psychological games. To start, we present a concrete example of a game in which common belief in rationality fails due to a peculiar discontinuity in players' utilities. Our example inspires a novel existence condition for belief hierarchies expressing common belief in rationality: *preservation of rationality at infinity*. Since all application-relevant psychological games do preserve rationality at infinity, our analysis shows that the possibility of common belief in rationality is not an issue in practice. This is different from the conventional existence condition *belief continuity* which imposes rather strict assumptions on players' utilities. To clarify this point, we present an example that shows how well-known departures from linear probability weighting like the certainty effect naturally lead to games that are not belief continuous but that do preserve rationality at infinity.

4.1. Impossibility of common belief in rationality

In traditional games, finiteness of player- and strategy-sets guarantees that common belief in rationality is possible. This is not true for psychological games, as the following example shows.[8,9]

Example 4.1 (*Common Belief in Rationality May not Be Possible*)**. Modified Bravery Game**: Player 1 chooses to act *timidly* or *boldly*, Player 2 is a passive observer. Player 1 is a timid guy, preferring to act timidly in almost all situations. But things change when Player 1 thinks that Player 2 considers his timidity a *commonly known fact*, not only believing that Player 1 chooses *timid*, but also believing that Player 1 believes that Player 2 believes that he chooses *timid*, and so on. Then Player 1 is angry and wants to act *boldly* to prove Player 2 wrong.

Let b_1^{timid} be the belief hierarchy for Player 1 where he believes that Player 2 believes it to be common knowledge that Player 1 is going to choose *timid*. So he believes that Player 2 believes that Player 1 chooses *timid*, believes that Player 2 believes that Player 1 believes that Player 2 believes that Player 1 chooses *timid*, and so on. Here, "believes" means "assigns probability 1 to".

Now let the Player 1's utility be such that $u_1(timid, b_1^{timid}) = 0$ and $u_1(bold, b_1^{timid}) = 1$, whereas $u_1(timid, b_1) = 1$ and $u_1(bold, b_1) = 0$ for every other belief hierarchy $b_1 \neq b_1^{timid}$. Hence, choice *bold* is rational for Player 1 iff his belief hierarchy is b_1^{timid} and *timid* is uniquely rational otherwise. The game is summarized in Table 1.

We now prove that there is no belief hierarchy for Player 1 that expresses common belief in rationality. We first show that the belief hierarchy b_1^{timid} does not express common belief in rationality. By definition, b_1^{timid} is such that Player 1 believes that

An alternative definition of static psychological games would require only condition (1) above. That definition would include games in which players move simultaneously in a single stage but where utilities potentially depend on *post-play* first- and higher-order beliefs regarding the choices of players. To model such psychological games, we might only need a static game form, but we *also* need the epistemic apparatus used to analyze dynamic games (i.e. a system of conditional belief hierarchies). For this reason, we find it more fruitful to draw the line between static psychological games and dynamic ones using requirements (1) *and* (2).

[8] The game is inspired by Geanakoplos et al.'s (1989) Bravery Game.

[9] For language-based games with the language $L_B(\Phi_\Gamma)$, a game that is similar to the one in Example 4.1 was independently developed by Bjorndahl et al. (2013). See Section 7 for details.

Table 1
Modified bravery game.

	$b_1 = b_1^{timid}$	$b_1 \neq b_1^{timid}$
timid	0	1
bold	1	0

Player 2 believes that Player 1 chooses *timid* and has belief hierarchy b_1^{timid}. But *timid* is not rational for the belief hierarchy b_1^{timid}, and hence under b_1^{timid}, Player 1 believes that Player 2 believes that Player 1 chooses irrationally. So b_1^{timid} does not express up to 2-fold belief in rationality and hence also not common belief in rationality.

Suppose, contrary to what we want to prove, that there exists a belief hierarchy b_1 for Player 1 that expresses common belief in rationality. Then, b_1 is such that Player 1 believes that Player 2 only assigns positive probability to belief hierarchies b_1' for Player 1 that express common belief in rationality. Since we have seen that b_1^{timid} does not express common belief in rationality, we conclude that b_1 must entail that Player 1 believes that Player 2 only assigns positive probability to belief hierarchies b_1' different from b_1^{timid}. Recall that only choice *timid* is rational for every such belief hierarchy b_1'. As under b_1, Player 1 must believe that Player 2 believes in Player 1's rationality, b_1 must imply that Player 1 believes that Player 2 believes that Player 1 chooses *timid*.

Moreover, b_1 must be such that Player 1 believes that Player 2 believes that Player 1 believes that Player 2 only assigns positive probability to belief hierarchies b_1' for Player 1 that express common belief in rationality. Hence, under b_1, Player 1 must believe that Player 2 believes that Player 1 believes that Player 2 only assigns positive probability to belief hierarchies b_1' different from b_1^{timid}. As choice *timid* is uniquely rational for every such belief hierarchy b_1', and b_1 is such that Player 1 believes that Player 2 believes that Player 1 believes that Player 2 believes in Player 1's rationality, it follows that, under b_1, Player 1 believes that Player 2 believes that Player 1 believes that Player 2 believes that Player 1 chooses *timid*.

Continuing in this fashion, we conclude that b_1 must be the belief hierarchy b_1^{timid} — a contradiction since we have seen that b_1^{timid} does not express common belief in rationality. Hence, there is no belief hierarchy for Player 1 that expresses common belief in rationality in this game.

In Example 4.1, common belief in rationality fails due to a peculiar discontinuity of Player 1's utility function. Player 1 strongly cares whether it is *commonly believed* that he will choose *timidly*: He strictly prefers *bold* iff he thinks it is *commonly believed* that he will act *timidly* and strictly prefers *timid* otherwise. An implication is that choice *timid* is not rational for the belief hierarchy b_1^{timid}, yet for every n there is a belief hierarchy $\hat{b}_1$ with $\hat{b}_1^n = (b_1^{timid})^n$ such that *timid* is rational for $\hat{b}_1$. As it turns out, this configuration is *typical* for games where common belief in rationality fails. If no player in a psychological game exhibits this discontinuous sensitivity to particular belief hierarchies, we say that the game *preserves rationality at infinity*. We now formally define this condition and prove that it is sufficient for common belief in rationality to be possible.

4.2. Preservation of rationality at infinity

The condition of *preservation of rationality at infinity* states that if a choice c_i is rational for every belief hierarchy in a sequence $b_i(1), b_i(2), \ldots$ where each $b_i(n-1)$ and $b_i(n)$ agree up to the $n-1$th-order belief, then c_i must also be rational for the corresponding limiting belief hierarchy.

Definition 4.2 (*Preservation of Rationality at Infinity*)**.** A psychological game Γ preserves rationality at infinity if the following is true for every player $i \in I$, every choice $c_i \in C_i$, and every belief hierarchy $b_i \in B_i$: Suppose that for every $n \geq 1$ there is some $\hat{b}_i \in B_i$ with $\hat{b}_i^n = b_i^n$ such that c_i is rational for $\hat{b}_i$. Then choice c_i is also rational for b_i.

Equivalently, whenever a choice c_i is *not* rational for a belief hierarchy b_i, then there must be some $n \geq 1$ such that c_i is not rational for any belief hierarchy $\hat{b}_i$ with $\hat{b}_i^n = b_i^n$.

We now prove that preservation of rationality at infinity ensures that common belief in rationality is possible. In the proof, we show how to *construct* a belief hierarchy expressing common belief in rationality under preservation of rationality at infinity. By contrast, all previous existence proofs (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009; Bjorndahl et al., 2013, etc.) proceed *non-constructively*.

Theorem 4.3 (*Possibility of Common Belief in Rationality*)**.** *Consider a psychological game Γ that preserves rationality at infinity. Then, for every player i, there exists a belief hierarchy $b_i \in B_i$ that expresses common belief in rationality.*

Comments on Theorem 4.3:

1. Our construction in the proof of Theorem 4.3 shows that, for any fixed $k \geq 1$, we can always find a belief hierarchy b_i for each player i that expresses up to k-fold belief in rationality, even in static psychological games that do not preserve rationality at infinity.[10] So belief in rationality can only ever fail when we try to extend a belief hierarchy expressing finitely many layers of up to k-fold belief in rationality to one that does so for all $k \in \mathbb{N}$.
2. In conjunction, Example 4.1 and Theorem 4.3 show that common belief in rationality only ever fails in psychological games where utility *discontinuously* depends on the *full* belief hierarchy. Hence, it is all but impossible to encounter this problem in any real-life application of psychological games. This fact could not have been inferred from previous existence results as in Geanakoplos et al. (1989), Battigalli and Dufwenberg (2009), etc. All previous results in the psychological games literature restrict to continuity of utility functions in the weak topology on B_i, an assumption that is easily violated already where utility depends only on finitely many levels of higher-order beliefs.[11] This explains why no concrete example of a game where common belief in rationality fails had been given in the psychological games literature prior to our Example 4.1.[9]
3. There also is a tight connection between the proof of Theorem 4.3 and Example 4.1: In the proof, we filter probability-one belief hierarchies, requiring ever higher levels of belief in rationality. Under preservation in rationality, we then find that the limit of the induced sequence of probability-one belief hierarchies must inherit all imposed rationality

[10] This continues to be true in games with infinite choice- and player-sets, provided that rational choices exist for any belief hierarchy b_i. See Footnote 33 for details.

[11] See Section 4.3 for examples. The existence condition for language-based games from Bjorndahl et al. (2013) is a notable exception among previously known results. Their condition is similar in flavor to ours in that a language-based utility function will pass the condition if, roughly speaking, rational and irrational states of mind can be distinguished based on finitely many logical formulae. However, as we argue in Section 7, Bjorndahl et al.'s (2013) result only applies for a subset of the games we consider here. And, moreover, it is not straightforward to check whether a given psychological game falls into the class of $L_B(\Phi_\Gamma)$-games they consider. And neither is it easy to verify for a given game within that class, whether utility functions satisfy Bjorndahl et al.'s (2013) sufficient condition for common belief in rationality or not.

requirements. Conversely, in Example 4.1 we put a similar filter on *all* belief hierarchies. Common belief in rationality is then seen to fail because increasing levels of belief in rationality single out a unique belief hierarchy that does *not* inherit all the imposed rationality requirements.

4. Preservation of rationality of infinity amounts to upper hemicontinuity of the best-response correspondence $BR_i(b_i) = \{c_i \in C_i |\ u_i(c_i, b_i) \geq u_i(c_i', b_i),\ \forall c_i' \in C_i\}$ in the product of discrete topologies on B_i. Based on this, we could also give a non-constructive proof of Theorem 4.3, applying Cantor's intersection theorem on the space of probability-one belief hierarchies.[12] Compared to this alternative, our constructive proof has the virtue of being completely elementary. Moreover, it nicely exposes the relationship between preservation of rationality at infinity and our Example 4.1 as described in comments (2) and (3) above.

4.3. Belief continuity

Based on Example 4.1 and Theorem 4.3, we now know that the impossibility of common belief in rationality is not a problem of empirical interest. A follow-up question is whether there are interesting psychological games that fail the existence conditions in previous papers but not ours. Here, we give a simple example of a psychological game where common belief in rationality is possible even though continuity as assumed in the previous literature fails to hold.

To start, we formally define the conventional continuity assumption, which we will call *belief continuity*. Geanakoplos et al. (1989) were the first to show that common belief in rationality is possible under this assumption.[13] In the following definition, $d(b_i^n, \hat{b}_i^n)$ denotes the Lévy–Prokhorov distance between the probability distributions induced by nth-order beliefs $b_i^n, \hat{b}_i^n \in B_i^n$.

Definition 4.4 (*Belief Continuity*).[14] A psychological game Γ is *belief-continuous* if, for every player $i \in I$, every choice $c_i \in C_i$, every belief hierarchy $b_i \in B_i$, and every $\varepsilon > 0$, there is $n \in \mathbb{N}$ and $\delta > 0$ such that for any belief hierarchy $\hat{b}_i \in B_i$ with $d(b_i^m, \hat{b}_i^m) < \delta$ for all $m \leq n$ we have that $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$.

Table 2
Modified bravery Game II.

	$b_1 \in B_1(\star, timid)$	$b_1 \notin B_1(\star, timid)$
timid	0	1
bold	1	0

Definitions 4.2 and 4.4 show that *belief continuity* is more restrictive than *preservation of rationality at infinity*: Preservation of rationality at infinity only asks that infinitely many levels of beliefs should never matter for the rationality of a choice-belief combination (c_i, b_i). Beyond that, belief continuity asks that utility functions vary continuously with trembles to the nth-order belief b_i^n for any $n \geq 1$. No restrictions of this second type are imposed under preservation of rationality at infinity. It is then easily seen that belief continuity implies preservation of rationality at infinity.[15]

Observation 4.5 (*Belief Continuity and Preservation of Rationality at Infinity*)**.** *Every belief-continuous game preserves rationality at infinity.*

More interestingly, it is easy to find games in which belief continuity fails but preservation of rationality at infinity does not. As the following example suggests, belief discontinuities naturally arise if we extend psychological game theory to account for non-linear probability weighting.[16] A similar example, *the indignant altruism game*, has been developed independently in Bjorndahl et al. (2013).

Example 4.6 (*Common Belief in Rationality without Belief Continuity*)**. Modified Bravery Game II**: Consider the following variation of Example 4.1: Player 1 now already gets angry if he believes that Player 2 *is sure* that Player 1 will act *timidly*. In that case, Player 1 wants to act *boldly* to prove Player 2 wrong.

Let $B_1(*, timid)$ be the set of Player 1's belief hierarchies where he believes that Player 2 believes that he chooses *timid*. Here, "believes" means "assigns probability 1 to". Player 1's utility function is given by $u_1(timid, b_1) = 1$, $u_1(bold, b_1) = 0$ for $b_1 \notin B_1(*, timid)$ and $u_1(timid, b_1) = 0$, $u_1(bold, b_1) = 1$ for $b_1 \in B_1(*, timid)$. So Player 1 strictly prefers *bold* iff he believes that Player 2 believes him to choose *timid* and strictly prefers *timid* otherwise. Table 2 summarizes the game.

It is easy to see that this game preserves rationality at infinity: Since utilities depend on at most second-order beliefs, c_i is necessarily rational for b_i whenever c_i is rational for some $\hat{b}_i$ with $\hat{b}_i^2 = b_i^2$. However, the game is not belief-continuous since slightly perturbing second-order beliefs for any $b_1 \in B_1(*, timid)$ leads to discontinuous changes in $u_1(bold, b_1)$. So while belief-continuity does not allow us to ascertain the possibility of common belief in rationality, preservation of rationality at infinity does. In fact, it is straightforward to find belief hierarchies that rationalize either choice for player 1 while expressing common belief in rationality.[17]

[12] The non-constructive proof is analogous to Theorem 5.5, Part 3. Specifically, we can use *preservation of rationality at infinity* to show, for all choices $c_i \in C_i$ and all $k \geq 1$, that the set of *probability-one* belief hierarchies that rationalize c_i under up to k-fold belief in rationality is closed in the space of *probability-one* belief hierarchies endowed with the product of discrete topologies. Now note that this space is compact by Tychonoff's theorem, and that up to k-fold belief in rationality can be satisfied for arbitrary finite k in any psychological game with finite choice sets C_i. It follows that there is a choice $c_i \in C_i$ such that the sets of probability-one belief hierarchies rationalizing c_i under up to k-fold belief in rationality define a sequence of nested non-empty compact sets. Using Cantor's intersection theorem, it follows that there exists a belief hierarchy that expresses common belief in rationality. One way to understand the filter on probability-one belief hierarchies that we apply in the constructive proof is as an explicit implementation of Cantor's intersection theorem. This explicit implementation heavily relies on the finiteness of choice sets C_i.

[13] More precisely, they prove that any belief-continuous static psychological game admits a psychological Nash equilibrium. Since psychological Nash equilibria are special instances of belief hierarchies expressing common belief in rationality, this implies that common belief in rationality is possible in all belief-continuous static psychological games. Battigalli and Dufwenberg (2009) prove that belief continuity makes common (strong) belief in rationality possible, not only in *static psychological games* as considered here but also in *dynamic psychological games*.

[14] This definition is slightly different from the one in Geanakoplos et al. (1989) in that we use Lévy–Prokhorov distances between nth-order beliefs whereas they use the composite metric that these Lévy–Prokhorov distances induce on B_i. It is easy to check that both definitions are equivalent (see Jagau and Perea, 2017 for a proof).

[15] For a simple proof, note that belief continuity amounts to continuity of utilities in the *product of weak topologies* on B_i. By contrast, preservation of rationality at infinity merely requires that each player i's best-response correspondence $BR_i(b_i) := \{c_i \in C_i | u_i(c_i, b_i) \geq u_i(c_i', b_i),\ \forall c_i' \in C_i\}$ be upper hemicontinuous in the stronger *product of discrete topologies* on B_i.

[16] While Example 4.6 is a belief-discontinuous *psychological* game, we could also produce violations of belief continuity in traditional settings. Take, for example, a Prisoner's dilemma game with the twist that one of the players strictly prefers to *cooperate* iff he is sure that his opponent will *defect*. That game will behave analogous to the game presented below, already without utilities that depend on higher-order beliefs.

[17] A helpful tool to find such belief hierarchies is the construction we used to filter belief hierarchies in the proof of Theorem 4.3. The reader may verify

The game from Example 4.6 might appear artificial, but it encapsulates a phenomenon that is well-known from other strands of the decision- and game-theory literature: In many experimental and real-life risky decisions, people are prone to the *certainty effect* (Tversky and Kahneman, 1981). Moving from almost certainty to certainty of an event can discontinuously change the evaluation of alternatives and thereby dramatically change behavior. Given the prevalence of the certainty effect in individual-decision settings, it is plausible that similar discontinuities can also play a role when agents reason about others' intentions and beliefs.

Clearly, whenever we want to model games in a way that allows for discontinuities in the processing of probabilities, we will automatically venture outside the class of *belief-continuous* games.[18] At the same time, already the fact that people in real-life decision problems plausibly care about at most finite levels of higher-order beliefs puts us squarely within the realm of games that *preserve rationality at infinity*.

5. Common belief in rationality characterized

In this section, we present a recursive elimination procedure that identifies the choices that can rationally be made under common belief in rationality in psychological games, similar to what *iterated elimination of strictly dominated choices* achieves in traditional games. Based on Definition 3.2 and Example 4.1 one might already expect that elimination among choices *and* beliefs will be needed to arrive at such a characterization. What is less apparent at this point is exactly which beliefs a potential elimination procedure must incorporate to work for a given game and a given set of belief-dependent utility functions. Our main result in this section, Theorem 5.5, gives a precise answer to this question: Whenever players' utilities depend on at most $n+1$th-order beliefs, elimination of choices and nth-order beliefs will be necessary and sufficient for a characterization. Leading up to Theorem 5.5, Section 5.1 presents a simple example to illustrate the precise role that belief elimination plays in our procedural characterization.

5.1. An introductory example

In the following, we present a simple psychological game involving guilt aversion. In the example, utilities depend on second-order beliefs, and it turns out that joint elimination of choices and first-order beliefs selects the choices that are consistent with common belief in rationality. As shown in Theorem 5.5, the example is representative of a more general pattern, leading to a characterization of common belief in rationality through elimination of choices and nth-order beliefs.

Example 5.1 (*Elimination of Choices and Beliefs in a Psychological Game*)**. A Date in the Pangs of Guilt**: You and Alice decided to have a date at a nice bar in town. Now it is the night of nights and you wonder whether to go to the *date* or to *stay* at home. At the other end of town, Alice is asking herself the same question.

Whether or not on a date, you enjoy going to bars. If it was not for the date, you would surely prefer not to stay home. However, you like Alice more as a friend than as a date. It is mostly the guilt you would feel from rejecting Alice that has made you agree to the date in the first place. Given this, you would actually prefer it if Alice just stayed home. Alternatively, you could stay home. Then however, you are in the pangs of guilt. Specifically, the more you believe that Alice goes to the date and expects you to come too, the worse you feel about staying home.

Different from a traditional game, your utility function therefore depends on both first- and second-order beliefs. Let it be defined as follows[19]:

$$u_y(date, b_y) = 1 - b_y^1(date),\ \ u_y(stay, b_y) = -\int_{\{date\}\times B_a} b_a^1(date)\ \mathrm{d}b_y$$

Here $\int_{\{date\}\times B_a} b_a^1(date)\ \mathrm{d}b_y$ represents the *expected probability* you assign to the event that Alice goes to the *date* and believes that you will come to the *date* too.[20]

Meanwhile, Alice prefers to go to the date if she thinks you will likely come, and otherwise she prefers to stay home. No different from a traditional game, Alice's utility function then only depends on first-order beliefs:

$$u_a(date, b_a) = b_a^1(date),\ \ u_a(stay, b_a) = 1 - b_a^1(date)$$

Already for this simple game, we can show that iterated elimination of strictly dominated choices will not suffice to characterize common belief in rationality. To see this, first note that every choice can be rationalized by at least one belief hierarchy for the respective player:

- For Alice, choosing *date* is rational whenever she believes that you choose *date* with probability greater than $\frac{1}{2}$ and *stay* is rational otherwise.
- For you, choosing *stay* is rational whenever you believe, with probability 1, that Alice chooses *date* and believes, again with probability 1, that you choose *stay*. Your choice *date*, on the other hand, is rational for any of your second-order beliefs.

Since every choice is rationalized by at least one belief for the respective player, it follows that iterated elimination of choices does not eliminate any choices for any player in this game.

However, we can easily show that both you and Alice can only choose *date* under common belief in rationality. To see this, it is sufficient to keep track of choices *and 1st-order beliefs*:

Step 1: You can only rationally choose *stay* if you put probability one on Alice's choice-1st-order-belief combination $(date, b_a^1)$ where $b_a^1(date) = 0$.
Alice can rationally choose *date* for any $b_a^1(date) \geq \frac{1}{2}$ and *stay* for any $b_a^1(date) \leq \frac{1}{2}$.

Step 2: If you believe in Alice's rationality, then you believe that she chooses *stay* if $b_a^1(date) \leq \frac{1}{2}$ and *date* if $b_a^1(date) \geq \frac{1}{2}$. But then, you cannot believe that Alice goes to the *date* for sure and believes $b_a^1(date) = 0$. Hence your choice *stay* cannot be rationalized for you under belief in Alice's rationality.

that $\hat{b}_i = b_i[(bold, *), (bold, *), (timid, *), (timid, *), (bold, *), (bold, *), \dots]$ is an example of a belief hierarchy that expresses common belief in rationality for both players i. Also, it can be shown that the game admits no psychological Nash equilibrium.

18 And even if we do not believe in such discontinuities in the processing of subjective probabilities, they might still be a useful approximation to a smoother underlying reality. As argued by Wakker (2010), stepwise-continuous probability-weighting functions in *Neo-Additive Rank-Dependent Utility* often strike an attractive balance between parsimony and fit when representing (inherently finite) experimental choice data in the decision-under-risk literature. Similar stepwise-continuous functions could be just as useful for modeling deviations from linear weighting of probabilities in experiments on belief-dependent motivations.

19 Your utility function captures a similar psychological mechanism as guilt aversion in Battigalli and Dufwenberg (2007). Different from that model, we consider a static game where utility depends on initial beliefs only. Such a variation of guilt aversion is not new, see Khalmetski et al. (2015).

20 A full fledged analysis of these *higher-order expectations* and *linear psychological games* is in Jagau and Perea (2021).

Step 3: If Alice entertains up to 2-fold belief in rationality, she must believe that you do not choose *stay*. Hence, *stay* cannot be rationalized for her under up to 2-fold belief in rationality.

It follows that only *date* is rational under common belief in rationality for both you and Alice.

Different from what we can observe in traditional games, there are no irrational choices for any player in Example 5.1, but there is a choice (your choice *stay*) that is not rational if you believe in Alice's rationality. By contrast, in a traditional game, there can be choices that are not rational under belief in the opponents' rationality only if there are irrational choices as well.

In Example 5.1, this happens because of an interaction between your preferences and Alice's rationality constraints[21]: Since you care not only about Alice's behavior but also about her (first-order) beliefs, knowing that Alice chooses rationally can directly matter for your preferences over actions. This is precisely the reason why iterated elimination of choices fails to characterize common belief in rationality in Example 5.1. In a traditional game, such interactions between preferences of players and rationality constraints of their opponents are excluded by construction. Players in traditional games only care about their opponents' *behavior*, so opponents' rationality constraints only have an *indirect impact* on preferences through excluding opponents' behaviors that turn out to be irrational at increasing levels of belief in rationality.

In Example 5.1, we still have an easy time figuring out the predictions under common belief in rationality, provided that we trace not only players' *choices* but also their *first-order beliefs*. It turns out that this intuition generalizes as follows: *Whenever utility depends on at most $n+1$th-order beliefs, common belief in rationality is characterized by iterative elimination of choices and nth-order beliefs.* We prove and discuss this result in the following section.

5.2. Iterated elimination of choices and nth-order beliefs

In traditional games, *iterated elimination of strictly dominated choices* selects exactly the choices that are consistent with common belief in rationality while only keeping track of players' choices. In this section, we generalize the result for traditional games by providing a procedure called *iterated elimination of choices and nth-order beliefs* that characterizes common belief in rationality in *belief-finite psychological games* while only keeping track of choices and finite levels of higher-order beliefs. We start by giving a definition of *belief-finite games*.

Definition 5.2 (*Belief-Finite Games*)**.** A psychological game Γ is belief-finite (of order n) if there is $n \geq 1$ such that for every player i, every choice $c_i \in C_i$, and every two belief hierarchies $b_i, \hat{b}_i \in B_i$ with $b_i^n = \hat{b}_i^n$ we have $u_i(c_i, b_i) = u_i(c_i, \hat{b}_i)$.

It is not hard to see that every belief-finite game preserves rationality at infinity.

Observation 5.3 (*Belief Finiteness and Preservation of Rationality at Infinity*)**.** *Every belief-finite game preserves rationality at infinity.*

Hence, by Theorem 4.3, common belief in rationality is possible in every belief-finite game.

[21] We can make the idea of such interactions precise using causality diagrams to model belief-dependent preferences. Mourmans (2019) uses this technique to characterize the classes of psychological games where iterated elimination of choices *does* suffice to characterize common belief in rationality.

Henceforth, we consider a belief-finite game of order $n+1$, so that we can write

$$u_i : C_i \times B_i^{n+1} \to \mathbb{R}.$$

We now show that *iterated elimination of choices and nth-order beliefs* exactly selects the choices consistent with common belief in rationality in belief-finite games of order $n+1$. To start, we define the procedure:

Procedure 5.4 (*Iterated Elimination of Choices and nth-Order Beliefs*)**.**

Step 1: For every player $i \in I$, define

$$\begin{aligned} R_i^n(1) = \{&(c_i, b_i^n) \in C_i \times B_i^n \mid \exists b_i^{n+1} \in B_i^{n+1} \\ &\text{with } \operatorname{marg}_{C_{-i} \times B_{-i}^{n-1}} b_i^{n+1} = b_i^n \\ &\text{such that } u_i(c_i, b_i^{n+1}) \geq u_i(c_i', b_i^{n+1}), \forall c_i' \in C_i\}. \end{aligned}$$

Step $k \geq 2$: Assume $R_i^n(k-1)$ is defined for every player i. Then, for every player i,

$$\begin{aligned} R_i^n(k) = \{&(c_i, b_i^n) \in R_i^n(k-1) \mid \exists b_i^{n+1} \in \Delta(R_{-i}^n(k-1)) \\ &\text{with } \operatorname{marg}_{C_{-i} \times B_{-i}^{n-1}} b_i^{n+1} = b_i^n \\ &\text{such that } u_i(c_i, b_i^{n+1}) \geq u_i(c_i', b_i^{n+1}), \forall c_i' \in C_i\}. \end{aligned}$$

We finally define: $R_i^n(\omega) = \bigcap_{k \in \{1,2,\dots\}} R_i^n(k)$, $\overline{R}_i^n = \bigcap_{k \in \text{Ord}} R_i^n(k)$.

Here Ord denotes the ordinals.[22] Transfinite elimination of (c_i, b_i^n) tuples will be necessary for similar reasons as transfinite elimination of non-best replies in Lipman (1994). See the third comment on Theorem 5.5 for details. We now prove:

Theorem 5.5 (*The Procedure Works*)**.** *Take a psychological game Γ that is belief-finite of order $n+1$.*

1. *For all $k \geq 0$, the choice-belief combinations $(c_i, b_i^n) \in C_i \times B_i^n$ that are consistent with up to k-fold belief in rationality are exactly the choice-belief combinations in $R_i^n(k+1)$.*
2. *The choice-belief combinations (c_i, b_i^n) that are consistent with common belief in rationality are exactly the choice-belief combinations in $\overline{R}_i^n$.*
3. *In a belief-continuous game, the choice-belief combinations (c_i, b_i^n) that are consistent with common belief in rationality are exactly the choice-belief combinations in $R_i^n(\omega)$.*

Comments on Theorem 5.5:

1. Elimination of choices and nth-order beliefs naturally generalizes the characterization of common belief in rationality in traditional games: Whenever utility depends on at most $n+1$th-order beliefs, we need to eliminate amongst choices and nth-order beliefs. So, in particular, if utility depends only on first-order beliefs ($n = 0$) our procedure turns into the familiar *iterated elimination of non-best replies* which is used to characterize correlated rationalizability in traditional games (Pearce, 1984; Tan and da Costa Werlang, 1988).[23]

[22] While any transfinite induction must terminate at *some* ordinal, the set of *all* ordinals is a paradoxical notion in Zermelo–Fraenkel set theory. So $k \in \text{Ord}$ should be interpreted as *iterating as needed* where the number of transfinite iterations of k cannot be further constrained other than that it is an ordinal. (That the output of transfinite iterations converges at *some* ordinal is guaranteed by the well-ordering theorem in conjunction with the fact that $\text{proj}_{C_i \times B_i^n} R_i(\omega)$ is a best-response set for any $n \geq 0$.) See Jagau (2021) for a more detailed investigation of transfinite eliminations of non-best replies.

[23] In Example 5.1, we actually do not need to track first-order beliefs for you since Alice's utility depended only on your choices. This suggests that we need

2. We can interpret Theorem 5.5 as an informational requirement for empirical investigations of psychological games. Since players' rationality in psychological games has to be judged in light of behavior *and* beliefs, psychological-games models cannot be studied relying only on behavioral data. Instead, we must combine information on subjects' behavior and beliefs to test the predictions of psychological game theory. This mirrors common practice in experiments on psychological games, where it is standard to elicit summary statistics of players' second-order beliefs.[24] Theorem 5.5 exposes the theoretical basis for these practices: If $n+1$th-order beliefs enter utility functions in our model, it is only ever identified based on data about subjects' choices and nth-order beliefs.
3. In belief-finite games of order $n+1$, the procedure always selects exactly the combinations of choices and nth-order beliefs that are consistent with *up k-fold-belief in rationality* for any finite k. However, if a game is not *belief-continuous*, it can happen that a tuple (c_i, b_i^n) is supported by a set of choices and belief hierarchies that shrinks towards the empty set as we iterate over up to k-fold belief in rationality. Such a (c_i, b_i^n) then survives ω steps of elimination of choices and nth-order beliefs without being consistent with *common belief in rationality*. This phenomenon is reminiscent of examples using traditional games with infinite choice sets (see (e.g.) Lipman, 1994; Dufwenberg and Stegeman, 2002; Bach and Cabessa, 2012), and it is easy to construct psychological-games counterparts of such examples. Different from traditional games, we do not need to assume infinite choice sets to generate such examples in belief-finite games of order $n > 1$. Since players' utility-relevant beliefs are distributions over $C_{-i} \times B_{-i}^n$, we can let the set of belief hierarchies b_i that rationalize some c_i shrink towards $\varnothing$ even if C_{-i} is finite. A concrete example is given in Appendix B.
   Similar to techniques used in infinite traditional games, we use transfinite iterations of our elimination procedure to deal with belief-discontinuous utility functions.[25] This shows an intriguing conceptual connection between belief-discontinuous psychological games and infinite traditional games that was not previously known.[26]

6. Extensions

6.1. Elimination of choices and nth-order beliefs and applications

Even though Theorem 5.5 considerably simplifies characterizing common belief in rationality relative to Definition 3.2, our procedure *iterated elimination of choices and nth-order beliefs* still fails to inherit many of the properties that have made *iterated elimination of strictly dominated choices* an attractive solution concept in traditional games. Clearly, this gap between psychological and traditional games could be closed under additional assumptions, and this would make the analysis of belief-dependent motivations in games significantly more manageable. Here, we sketch three particularly promising ways in which one could hope to simplify psychological games:[27]

- **Linearity:** In traditional games, dependence of utilities on first-order beliefs and the expected-utility assumption turn *iterated elimination of non-best replies* into a sequence of linear programs. By contrast, we face two sources of non-linearity in psychological games: Firstly, u_i may depend non-linearly on b_i. Secondly, b_i itself is a highly non-linear object. To restore linearity, we would need to extend the expected-utility assumption to psychological games.[28]
- **Finiteness:** In traditional games, *iterated elimination of non-best replies* has converged as soon as we are not able to eliminate any choices for any player for the first time since the start of the procedure. Also, at any given step $k > 1$, we will be able to eliminate choices for a player i iff choices for *some* of i's opponents were eliminated at step $k-1$. This immediately implies that $\sum_{i \in I} |C_i| - 1$ is an upper bound on the length of the procedure in traditional games. By contrast, in psychological games it is *typically* the case that no upper bound on elimination of choices and nth-order beliefs can be given, and we can construct examples of simple psychological games where infinitely many eliminations of choices and nth-order beliefs occur.[29] Intuitively, the reason is that multilaterally belief-dependent preferences can lead to a feedback loop of overlapping restrictions, where rationality constraints of different players interact to restrict admissible beliefs further and further ad infinitum. This cannot happen in (finite) traditional games, where opponents' rationality constraints exhaust their influence on a given player i's choice problem through excluding certain behaviors of i's opponents.[30] Different from the previously discussed non-linearity of psychological utility, these feedback loops should be considered a feature rather than a bug in most interesting applications.
- **Pearce's Lemma:** In traditional games, Pearce's Lemma (Pearce, 1984) allows us to replace *iterated elimination of non-best replies* with *iterated elimination of strictly dominated choices*. This greatly reduces the dimensionality of the underlying utility maximization problem, allowing us to limit comparisons of state-dependent utility among choices $c_i \in C_i$ to all states in C_{-i} rather than having to consider every probabilistic state in $\Delta(C_{-i})$. Pearce's Lemma fails in psychological games for much the same reasons as the linearity of *elimination of non-best replies*. We may expect to restore it under qualitatively similar (if stronger) assumptions.

only track beliefs for a player up to the level that they enter *some* opponent's utility. Indeed, asymmetric situations of this sort can be covered by extending Theorem 5.5 as follows: *Let Γ be a belief-finite, belief-continuous game where $n_1 \geq n_2 \geq \cdots \geq n_{|I|} \geq 1$ are the highest utility-relevant levels of higher-order beliefs for players $1, \ldots, |I|$. Then common belief in rationality is characterized by iterated elimination of choices and $n_1 - 1$th-order beliefs for players $2, \ldots, |I|$ and of choices and $n_2 - 1$th-order beliefs for player 1.*

24 In Jagau and Perea (2021), we show that such summary statistics are not only convenient for empirical investigations, but that we can construct a whole simplified version of psychological game theory where belief hierarchies are replaced by suitably defined summary statistics as the new epistemic primitive.

25 This should not be understood as a refinement of common belief in rationality. Rather, all steps $k \geq \omega$ of transfinite elimination of choices and nth-order beliefs can be equated to step ω in Definition 3.2. See Jagau (2021) for details.

26 The connection to infinite traditional games is further discussed in Section 6.2.

27 See Jagau and Perea (2021) for an in-depth investigation of finite and linear procedures and applications of Pearce's Lemma in psychological games.

28 In Jagau and Perea (2021), we show that psychological utility functions can be fully linearized by iterating expected-utility-type assumptions across all levels of higher-order beliefs. Your utility function in Example 5.1 is an instance of such a linear psychological (expected) utility function.

29 The example in Appendix B involves a simple game between you and Alice in which utilities only depend on second-order beliefs, and where we require ω-fold elimination of choices and first-order beliefs to select the tuples of choices and first-order beliefs that are consistent with common belief in rationality.

30 For infinite traditional games, a similar mechanism turns out to cause *transfinite* iterations of eliminations of non-best replies. See Jagau (2021) for details.

6.2. Infinite choice- and player-sets

Theorem 4.3, Example 4.1, and Theorem 5.5 show a striking similarity between strategic reasoning in *psychological games* and strategic reasoning in *traditional games with infinite choice and/or player sets*: Firstly, in both classes of games, it is not guaranteed that common belief in rationality can be satisfied, even if arbitrary finite orders of belief in rationality are satisfiable (see Dufwenberg and Stegeman, 2002, Example 2). Secondly, selecting the choices consistent with common belief in rationality might necessitate an elimination procedure of *arbitrary transfinite length*, rather than a countably infinite one (see Lipman, 1994).

Mathematically, the reason for this similarity lies in the richness of the utility-relevant state space: In infinite traditional games, this richness is immediate in the infinite size of the set of opponents and/or the sets of opponents' choices. In psychological games we get the same richness for any (non-singleton) size of player- and choice-sets, since, for any $n > 1$, player i's nth-order belief b_i^n is a probability distribution on the uncountably infinite set $C_{-i} \times B_{-i}^{n-1}$. Here as there, rich state spaces start to be a problem whenever utility is *discontinuous* on the respective state space. And, as Theorem 5.5, Part 3 shows, restoring this continuity in psychological games has the same effects as in infinite traditional games (i.e. a non-empty maximal reduction exists and can be characterized through standard induction; Dufwenberg and Stegeman, 2002, Theorem 1).

A notable implication of these remarks is that the gap between finite psychological games as in Definition 2.1 and psychological games with infinite choice- and player-sets becomes very small:

Observation 6.1 (*Common Belief in Rationality in Infinite Psychological Games*)**.** *Consider a static psychological game* $\Gamma = (C_i, B_i, u_i)_{i \in I}$ *with* I *a countable set of players,* C_i *a separable set of choices for player* i*,* B_i *the set of belief hierarchies for player* i *expressing coherency and common belief in coherency,*[31] *and* $u_i : C_i \times B_i \to \mathbb{R}$ *a measurable and bounded utility function.*

Possibility:

1. *For every player* $i \in I$*, let* u_i *be such that, for every* $b_i \in B_i$*, some* $c_i \in C_i$ *is rational given* b_i*.*[32] *Then, for every player* i *and* $k \geq 1$*, there exists a choice-belief hierarchy combination* (c_i, b_i) *expressing up to* k*-fold belief in rationality.*[33]
2. *If, in addition,* C_i *is compact Hausdorff for each player* $i \in I$ *and* Γ *is belief-continuous, then there exists a choice-belief hierarchy tuple* (c_i, b_i) *expressing rationality and common belief in rationality for every player* i*.*

Procedural Characterization:

1. *If* Γ *is belief-finite of order* $n + 1$*, then a choice-belief tuple* (c_i, b_i^n) *is consistent with common belief in rationality iff it survives transfinite elimination of choices and* n*th-order beliefs.*
2. *If, in addition,* C_i *is compact Hausdorff for each player* $i \in I$ *and* Γ *is belief-continuous, then a choice-belief tuple* (c_i, b_i^n) *is consistent with common belief in rationality iff it survives* ω*-fold elimination of choices and* n*th-order beliefs.*

Observation 6.1 and the previous remarks show that modeling challenges around infinite psychological games are essentially identical to what we encounter in infinite traditional games. Also, it is worth noting that the weakest set of assumptions we have considered here (measurability of u_i and separability of C_i for all players $i \in I$) exactly coincides with minimal assumptions that are needed to construct a universal type space for any given game (Heifetz and Samet, 1998, 1999) and to ensure that rationality is a well-defined event within that type space. As such, these assumptions can be regarded as a prerequisite for any form of psychologically game-theoretic analysis.

6.3. Dynamic psychological games and asymmetric information

Our results in this paper are restricted to static psychological games as originally defined in Geanakoplos et al. (1989) (see Definition 2.1). Other applied and theoretical literature covers the richer class of dynamic psychological games in which players move sequentially and utility is allowed to depend on updated beliefs (see Battigalli and Dufwenberg, 2009), as well as psychological games under incomplete information (see Attanasi et al., 2016, Battigalli et al., 2020) and psychological games under unawareness (see Sebald, 2010).

For brevity, we do not explicitly extend our analysis to these broader classes of games in the present paper. However, there are strong grounds to expect our results would extend to sequential interaction and asymmetric information very smoothly.

Common belief in rationality in static games has a direct dynamic counterpart in *common belief in initial rationality* (Ben-Porath, 1997). Extended in this fashion, common belief in rationality does not restrict how players update beliefs as a dynamic game unfolds — different from stronger reasoning concepts such as sequential rationalizability and common belief in future rationality (Dekel et al., 1999, 2002; Asheim and Perea, 2005; Perea, 2014) and extensive form rationalizability and common strong belief in rationality (Pearce, 1984; Battigalli, 1997; Battigalli and Siniscalchi, 2002). In dynamic psychological games, many novel and interesting phenomena arise precisely because of how consistency restrictions across initial and updated conditional belief hierarchies cause players to *revise* their understanding of opponents' states of mind in light of the history of play. As demonstrated in Battigalli et al. (2020), this leads to additional challenges when we extend reasoning concepts involving restrictions on belief revision (like common strong belief in rationality or common belief in future rationality) from dynamic traditional games to dynamic psychological games. However, like in the case of static psychological games, allowing belief-dependent utility

[31] Whenever C_i is not Polish, we need to refine Brandenburger and Dekel's (1993) construction of B_i. Specifically, if choice sets are assumed to be general measurable spaces, it is not necessarily the case that a hierarchy of coherent beliefs $b_i^1, b_i^2, \ldots$ induces a probability distribution on opponents' choices and belief hierarchies (see Heifetz and Samet, 1999). Instead, $b_i^1, b_i^2, \ldots$ might distribute on a decreasing sequence of sets $X_i^n \subset C_{-i} \times B_{-i}^{n-1}$ such that ultimately no σ-additive measure on the space of belief hierarchies can be induced by the sequence of finite-order beliefs. The obvious solution is to require the existence of such a σ-additive measure in our definition of coherent belief hierarchies. Heifetz and Samet's (1998,1999) results imply that adding this requirement to Brandenburger and Dekel's (1993) construction indeed defines a space of belief hierarchies B_i such that each $b_i \in B_i$ satisfies coherency and common belief in coherency and can be identified with a unique probability measure $\delta(b_i) \in \Delta(C_{-i} \times B_{-i})$.

[32] For infinite traditional games, Theorem 1 in Dufwenberg and Stegeman (2002) instead requires $u_i(c_i, b_i^1)$ to be upper semi-continuous on C_i for each $b_i^1 \in B_i^1$. Clearly, rational choices will exist for any belief hierarchy if u_i is upper semi-continuous in this sense, whereas the reverse implication is not true.

[33] Whenever rational choices can fail to exist for some beliefs b_i of some player i, rationality not only eliminates irrational choices but also beliefs for which no rational choice exists. At any level $k \geq 0$ of rationality and up to k-fold belief in rationality, it might now happen that *all* surviving beliefs for some player i admit no rational choice. In that case, rationality and up to k-fold belief in rationality fails for i, and rationality and up to $k + m$-fold belief in rationality fails for all $m \geq 1$ and all players $j \in I$. Elimination procedures and definitions continue to work regardless as long as they account for the new belief-elimination part of the rationality constraint throughout. Our Definition 3.2 and Procedure 5.4 achieve this by reimposing the original rationality constraint on every elimination step. This approach is not new, see Bernheim (1984), Milgrom and Roberts (1990), and Chen et al. (2007).

into dynamic such games *in the first place* must also bring in the more basic complications for strategic reasoning that we have explored here. These issues are logically prior to considerations around belief revision and consistency of strategic rationality and belief-dependent utility across time. Hence, it is all but inevitable that our results constrain any analysis of dynamic psychological games in essentially the same fashion as they constrain the analysis of static psychological games.[34]

Essentially the same reasoning applies relating to richer models of uncertainty and subjective perceptions of the game environment (as under incomplete information and unawareness). Surely, this introduces novel challenges for players to *interpret* opponents' moves in light of their belief-dependent preferences, and these will be borne out by constraints on suitable extensions of common belief in rationality. However, also here the problem of reconciling belief-dependent utility with strategic reasoning in general must logically precede any additional phenomena that might arise.

Obviously, this does not mean that studying the interaction of belief-dependent motivations with a time structure (as in dynamic psychological games), exogenous sources of uncertainty (as incomplete-information games), and differing subjective models of the interaction (as in games with unawareness) is not interesting and important in its own right. Clearly, our remarks here cannot stand in for a full-fledged extension of Theorems 4.3 and 5.5 and Observation 6.1 to models of strategic rationality in dynamic and/or asymmetric-information psychological games.[35] We believe that such an extension would require extensive foundational work,[36] and – as such – it would make for an interesting follow-up to the analysis presented here.

7. Related literature

7.1. Psychological game theory

In this paper we have focused on *common belief in rationality* in static psychological games.

Our first concern was to determine sufficient conditions for common belief in rationality to be possible. Previously Kolpin (1992), Battigalli and Dufwenberg (2009), Sebald (2010), and Battigalli et al. (2020) have provided existence proofs for rationalizability and equilibrium in various settings.[37] Different from our Theorem 4.3, all of these results essentially rely on the same continuity condition as Geanakoplos et al. (1989) use in their original paper. Furthermore, none of the mentioned contributions provides an explicit example of a psychological game in which rationalizable strategies fail to exist. We present such a game in our Example 4.1.

Our second goal was to develop a procedural characterization for common belief in rationality in psychological games. Battigalli et al. (2020) provide a characterization of common strong belief in rationality in incomplete-information dynamic psychological games, relying on measure-theoretic techniques and continuity assumptions analogous to Battigalli and Dufwenberg (2009).[38] We present an alternative characterization that relies on Brandenburger and Dekel's (1993) construction of belief hierarchies expressing coherency and common belief in coherency. In virtue of this, our approach more readily connects to standard epistemic game theory. In addition, our characterization generalizes to the case of discontinuous utility functions and infinite choice sets — an extension that had not previously been explored either in Battigalli et al. (2020) or elsewhere in the psychological-games literature.

7.2. Language-based games

Bjorndahl et al. (2013) introduce language-based games in which, given a static game form $G = (C_i)_{i \in I}$, a probabilistic Kripke frame describing players' beliefs over a space of states of the game Ω, and a language L collecting logical propositions about the state of the game, utility functions of players are defined on so-called *L-situations* $S(L)$ — maximal satisfiable sets of formulae within the language L.

Their leading example of a "non-standard" language, denoted $L_B(\Phi_\Gamma)$, allows a player i to express propositions not only about opponents' choices C_{-i}, but also about which choices C_{-j} an opponent j deems possible for their opponents, which choices C_{-k} an opponent j deems possible that their opponent k deems possible for their opponents, and so on. As such, $L_B(\Phi_\Gamma)$-games look very similar to static psychological games as in our Definition 2.1 and in the Observation 6.1-generalization. In fact, a closer inspection suggests that $L_B(\Phi_\Gamma)$-games correspond to a subset of the psychological games we consider. To see this, note that, while the Kripke frame modeling players' differential information regarding states of the game Ω is probabilistic, the language $L_B(\Phi_\Gamma)$ only allows players to care about whether a vector of opponents' choices c_{-i} is deemed possible at some state $\omega \in \Omega$, about whether it is possible at ω that an opponent j deems possible a vector of opponents' choices c_{-j}, and so on. Continuing in this way, is then not hard to see that the $L_B(\Phi_\Gamma)$-situations that players may care about in Bjorndahl et al. (2013) closely correspond to sets of *probability-one* belief hierarchies in our paper. By contrast, players in our paper are allowed to care in arbitrary ways about *probabilistic* beliefs b_i^k of arbitrary order $k \geq 1$, up to and including the full *probabilistic* belief hierarchy b_i.[39] In fact, it is not hard to find psychological games that fall under our Definition 2.1 but cannot be modeled as an $L_B(\Phi_\Gamma)$-game. The game in Appendix B is one such example.[40]

[34] Maybe the clearest sign of this strong interdependence is Battigalli and Dufwenberg's (2009) and Battigalli et al.'s (2020) reliance on generalized versions of Geanakoplos et al.'s (1989) belief continuity as fundamental assumptions in their respective analyses.

[35] A sketch of how definitions and results translate to dynamic games and asymmetric information is available from the authors upon request.

[36] This would concern, in particular, the results presented in Section 6.2. For that part of our analysis, we rely on methods to construct a universal type space that have not been extended to the case of dynamic games. See Footnote 31 for details.

[37] Early models in Geanakoplos et al. (1989) and Kolpin (1992) restrict to equilibrium analysis in static and sequential psychological games where utility functions are assumed to depend on initial beliefs only. Battigalli and Dufwenberg (2009) also continue this project by studying sequential equilibrium (Kreps and Wilson, 1982) in their setup.

[38] Battigalli et al. (2020) start out with general (coherent and incoherent) belief hierarchies and then proceed to iteratively impose coherency, rationality and belief in coherency and rationality at increasing levels. That alternative way of proceeding leads to the same results as ours as far as belief-continuous games are concerned.

[39] Furthermore, Bjorndahl et al. (2013) assume that players use expected utility given their subjective belief at each state $\omega \in \Omega$ to weigh utilities derived from deterministic situations. Jagau and Perea (2021) study a class of *linear* psychological games. The characteristic assumption on utilities in these games, *linearity in level-k uncertainty for all* $k \geq 1$, is similar to (albeit slightly more restrictive than) the linearity- and invariance-assumptions characterizing utility in $L_B(\Phi_\Gamma)$-games.

[40] To see this, consider Bob's utility, given by Table 4. Bob's utility u_b assigns 0 to choosing *CBR* whenever $\text{Supp}(b^2_{b,a})$ contains some B, b^a_1 with $b^a_1(B) < \frac{1}{\sqrt{3}}$ and 1 to choosing *CBR* given any $b^2_{b,a}$ that does not have this property. Clearly, there is no assignment of *CBR*-utility values to finite formulas about choices and probability-one first-order beliefs of Alice (i.e. all sets comprised of (B, B), (B, F), (F, B), (F, F), and the projections of these tuples onto either of their coordinates) that could capture this discontinuity of u_b on $B^2_{b,a}$. Hence the game from Appendix B cannot be modeled as an $L_B(\Phi_\Gamma)$-game.

Within the narrower class of $L_B(\Phi_\Gamma)$-games, Bjorndahl et al. (2013) present the "deeply surprising proposal game" to demonstrate that rationalizable choices may not exist without further restrictions. This example is similar to the modified bravery game (see Example 4.1), which we use to show that common belief in rationality might be impossible in general psychological games.

Their non-existence example leads (Bjorndahl et al., 2013) to introduce the condition *CR*, which states that, whenever a player chooses irrationally at an $L_B(\Phi_\Gamma)$-situation, it should be possible to conclude this fact already on the basis of *finitely many formulae*. This may be viewed as a *logical compactness* requirement on the set of $L_B(\Phi_\Gamma)$-statements regarding a given game. Accordingly, *CR* allows (Bjorndahl et al., 2013) to show that the set of $L_B(\Phi_\Gamma)$-situations where all players choose rationally is *closed* within $S(L_B(\Phi_\Gamma))$, from whence the existence of rationalizable choices then follows using standard topological arguments.

In this paper, we also introduce a sufficient condition for the possibility of common belief in rationality (or, equivalently, for the existence of rationalizable choices). Our condition, *preservation of rationality at infinity* (see Definition 4.2), states that, whenever a choice is not optimal for a belief hierarchy b_i, there must be some level n such that the choice is not optimal for any belief hierarchy that coincides with b_i at the first n layers. In other words, knowing *finitely many* layers of belief is sufficient for concluding that a choice is sub-optimal for a given belief hierarchy. In that sense, *preservation of rationality at infinity* is similar in spirit to Bjorndahl et al.'s (2013) *CR*.

Different from *CR*, our *preservation of rationality at infinity* is a non-topological condition, which is directly verifiable for any given tuple of a choice and a belief hierarchy. And importantly, also our proof of Theorem 4.3 is completely elementary and constructive: In the proof, we show how a belief hierarchy that expresses common belief in rationality can be constructed for every player under our structural assumptions. By contrast, Bjorndahl et al. (2013) present a non-constructive proof, relying on a version of Cantor's intersection theorem. Hence, in particular, their proof does not provide a method to explicitly construct $L_B(\Phi_\Gamma)$-situations that express common belief in rationality in a given game.

Furthermore, while a non-constructive proof for Theorem 4.3 could be given as well (see the comments following the theorem), that proof uses more structure on the space of choice-belief hierarchy tuples $C_i \times B_i$ than what is furnished by Bjorndahl et al.'s (2013) assumptions, were we to directly translate them to our setting. Specifically, results from Heifetz and Mongin (2001) and Meier (2012) imply that an infinitary language L would generally be needed to model psychological games as in Definition 2.1 using the language-based approach.

8. Conclusion

Since its introduction by Geanakoplos et al. (1989), psychological game theory has become a popular tool to formally capture numerous belief-dependent motivations and their role in strategic interaction. Nevertheless, our theoretical understanding of psychological games still falls short of what we are used to from traditional games. In this paper we have provided an in-depth analysis of how belief-dependent utility interacts with common belief in rationality, the basic building block in all models of strategic rationality. Our analysis zooms in on the elemental setting of static psychological games, and it focuses on two basic questions:

1. What minimal assumptions make common belief in rationality possible?
2. How can we characterize common belief in rationality using a recursive elimination procedure?

Regarding question (1), we introduce a new existence condition, *preservation of rationality at infinity*. This allows us to show that common belief in rationality is possible in any empirically-relevant psychological game. This was far from obvious based on previous existence results (Geanakoplos et al., 1989, Battigalli and Dufwenberg, 2009, etc.) that rely on significantly stronger assumptions.

Regarding question (2), we prove that (possibly transfinite) *iterated elimination of choices and nth-order beliefs* characterizes common belief in rationality in all situations where players' utilities depend on at most $n+1$*th-order beliefs*. Our characterization extends previously known ones in that it relies on minimal assumptions regarding the functional form of belief-dependent utilities.

Building on our main analysis, we also extend our results to psychological games with infinite choice- and player-sets, a class of games which has rarely, if ever, been investigated before.

Our results regarding the possibility and characterization of common belief in rationality in psychological games dramatically relax a set of long-standing assumptions that have been shared by all previous models in psychological game theory up to and including Geanakoplos et al.'s (1989) seminal paper. Moreover, they uncover interesting and previously unknown parallels between models of strategic reasoning in psychological games and (infinite) traditional games. As such, they are best suited to demonstrate a resting need to return to simple cases of psychological games in order to sharpen our intuitions concerning the mechanics of belief-dependent utility. An extension of the present analysis to richer models of psychological games involving sequential interaction and asymmetric information is an interesting avenue for future research.

Table 3
Two-tiered battle of the sexes.

	b_y^1	
You	B	F
B	2	0
F	0	1

+

	e_y^2	
You	$(\cdot, B)$	$(\cdot, F)$
B	2	0
F	0	1

	b_a^1	
Alice	B	F
B	1	0
F	0	2

+

	e_a^2	
Alice	$(\cdot, B)$	$(\cdot, F)$
B	1	0
F	0	2

Table 4
Outside wager.

Bob	$b_{b,a}^2 \in \Delta(\overline{R}_a^1)$	$b_{b,a}^2 \notin \Delta(\overline{R}_a^1)$
CBR	1	0
$\neg$*CBR*	0	1

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

This example also helps illustrate why an infinitary language L would really be needed to capture the full scope of belief-dependent utility that our Definition 2.1 allows for. In particular, adding one intermediate degree of belief to $L_B(\Phi_\Gamma)$ (analogous to Example 3.2 in Bjorndahl et al., 2013) would allow us to capture games as the one in Appendix B with a fixed cutoff k such that Alice's choice-belief tuples with $c_a = B$ are consistent with common belief in rationality iff $b_1(B) \geq k$. However, our Definition 2.1 implies a language that allows us to distinguish Appendix B-type games as we continuously vary k over $\mathbb{R}$ by perturbing the utilities for you and Alice given in Table 3. Clearly, there is no finitary generalization of $L_B(\Phi_\Gamma)$ that can distinguish between all of those games at the same time.

Acknowledgments

This paper benefited from the suggestions of audiences at Claremont Graduate University, Maastricht University, Northwestern University, the University of California, Irvine, the Workshop on Psychological Game Theory 2017, the TARK Conference 2017, the SABE/IAREP Conference 2018, the 33rd Annual Congress of the European Economic Association, and the Southwestern Economic Theory Conference 2019. We thank Pierpaolo Battigalli, Andrés Carvajal, Jean-Paul Carvalho, Martin Dufwenberg, János Flesch, Donald G. Saari, Marciano Siniscalchi, and two anonymous referees for comments. Stephan Jagau gratefully acknowledges financial support from the Netherlands Organisation for Scientific Research (NWO) under Rubicon Grant 19.181SG.023.

Appendix A. Proofs

Proof of Theorem 4.3. For the proof we need a new piece of notation. Consider, for every $n \geq 1$, a choice profile $c^n = (c_i^n)_{i \in I}$ in $\bigtimes_{i \in I} C_i$. Then, we denote by $b_i[c^1, c^2, \ldots]$ the belief hierarchy for player i that (1) for every $j \neq i$, assigns probability 1 to choice c_j^1, (2) for every $j \neq i$ and every $k \neq j$, assigns probability 1 to the event that j assigns probability 1 to choice c_k^2, and so on. As an abbreviation, we denote the nth order belief of $b_i[c^1, c^2, \ldots]$ by $(c^1, \ldots, c^n)$, and thus write $b_i^n[c^1, c^2, \ldots] = (c^1, \ldots, c^n)$.

We will now generate, for all players i, an infinite set of belief hierarchies

$$\hat{B}_i = \{b_i(0), b_i(1), b_i(2), \ldots\}$$

as follows. Select, for every $n \geq 1$, an arbitrary choice profile $c^n = (c_i^n)_{i \in I}$ in $\times_{i \in I} C_i$ and set

$$b_i(0) := b_i[c^1, c^2, \ldots]$$

for every player i. Moreover, for every player i let $d_i(1)$ be a choice that is rational for $b_i(0)$, and set $d(1) := (d_i(1))_{i \in I}$. Then, for all players i, define a new belief hierarchy

$$b_i(1) := b_i[d(1), c^1, c^2, \ldots]$$

and let $d_i(2)$ be a choice that is rational for $b_i(1)$. Set $d(2) := (d_i(2))_{i \in I}$. Subsequently, for all players i, define the new belief hierarchy

$$b_i(2) := b_i[d(2), d(1), c^1, c^2, \ldots],$$

and so on. By construction, the belief hierarchy $b_i(n) \in \hat{B}_i$ expresses up to n-fold belief in rationality, for every player i and every $n \geq 1$.

We now construct, for a given player i, a belief hierarchy $\hat{b}_i$, as follows. Since there are only finitely many choices, there is a choice profile $e^1 = (e_j^1)_{j \in I}$ in $\bigtimes_{j \in I} C_j$ such that there are infinitely many belief hierarchies $b_i \in \hat{B}_i$ with $b_i^1 = e^1$. Let

$$\hat{B}_i[e^1] := \{b_i \in \hat{B}_i | b_i^1 = e^1\},$$

which is an infinite set, by construction. But then, there must be a choice profile $e^2 = (e_j^2)_{j \in I}$ in $\bigtimes_{j \in I} C_j$ such that there are infinitely many belief hierarchies $b_i \in \hat{B}_i[e^1]$ with $b_i^2 = (e^1, e^2)$. Let

$$\hat{B}_i[e^1, e^2] := \{b_i \in \hat{B}_i | b_i^2 = (e^1, e^2)\},$$

which again is an infinite set, by construction. Hence, there must be a choice profile $e^3 = (e_j^3)_{j \in I}$ in $\bigtimes_{j \in I} C_j$ such that $b_i^3 = (e^1, e^2, e^3)$ for infinitely many belief hierarchies $b_i \in \hat{B}_i[e^1, e^2]$. Let

$$\hat{B}_i[e^1, e^2, e^3] := \{b_i \in \hat{B}_i | b_i^3 = (e^1, e^2, e^3)\},$$

which again is an infinite set, by construction. By continuing in this fashion, we obtain an infinite sequence of choice profiles $e^1, e^2, \ldots$, and we set

$$\hat{b}_i := b_i[e^1, e^2, \ldots].$$

We now show that $\hat{b}_i$ expresses common belief in rationality. That is, we must show, for every $n \geq 1$ and every player j, that choice e_j^n is rational for the belief hierarchy $b_j[e^{n+1}, e^{n+2}, \ldots]$. Fix such an n and a player j.

Since the game preserves rationality at infinity, it suffices to show that for every $m \geq 1$ there is some $b_j \in B_j$ with $b_j^m = b_j^m[e^{n+1}, e^{n+2}, \ldots]$ such that e_j^n is rational for b_j. Let $m \geq 1$ be fixed. Since $\hat{B}_i[e^1, \ldots, e^{n+m}]$ is an infinite subset of $\hat{B}_i$, there is $k \geq n$ such that $b_i(k) \in \hat{B}_i[e^1, \ldots, e^{n+m}]$. Let

$$b_i(k) = b_i[e^1, \ldots, e^{n+m}, g^{n+m+1}, g^{n+m+2}, \ldots],$$

where $g^{n+m+1}, g^{n+m+2}, \ldots$ are choice profiles in $\bigtimes_{i \in I} C_i$.

Define the belief hierarchy

$$b_j := b_j[e^{n+1}, \ldots, e^{n+m}, g^{n+m+1}, g^{n+m+2}, \ldots].$$

Then, by construction, $b_j^m = (e^{n+1}, \ldots, e^{n+m}) = b_j^m[e^{n+1}, e^{n+2}, \ldots]$. Moreover, since $b_i(k)$ expresses up to k-fold belief in rationality, and $k \geq n$, we conclude that $b_i(k)$ expresses up to n-fold belief in rationality. Since $b_i(k) = b_i[e^1, \ldots, e^{n+m}, g^{n+m+1}, g^{n+m+2}, \ldots]$, it follows that e_j^n is rational for $b_j[e^{n+1}, \ldots, e^{n+m}, g^{n+m+1}, g^{n+m+2}, \ldots] = b_j$. Hence, for every $m \geq 1$ we can construct in this fashion some $b_j \in B_j$ with $b_j^m = b_j^m[e^{n+1}, e^{n+2}, \ldots]$ such that e_j^n is rational for b_j. As the game preserves rationality at infinity, we conclude that e_j^n is rational for the belief hierarchy $b_j[e^{n+1}, e^{n+2}, \ldots]$. Since this holds for every $n \geq 1$ and every player j, the belief hierarchy $\hat{b}_i := b_i[e^1, e^2, \ldots]$ expresses common belief in rationality.

Therefore, in this fashion we can construct for every player i a belief hierarchy $\hat{b}_i$ that expresses common belief in rationality. This completes the proof. □

Proof of Theorem 5.5.
Part 1:

⇒ We start by showing that any (c_i, b_i^n) that is consistent with up to k-fold belief in rationality is in $R_i^n(k+1)$. We proceed by induction over $k \geq 0$.

Induction Start: Let (c_i, b_i^n) be consistent with 0-fold belief in rationality. Then c_i is rational for a belief hierarchy b_i that induces b_i^n. So, in particular, the $n+1$th-order belief b_i^{n+1} induced by b_i must satisfy $u_i(c_i, b_i^{n+1}) \geq u_i(c_i', b_i^{n+1})$, $\forall c_i' \in C_i$. It follows that $(c_i, b_i^n) \in R_i^n(1)$.

Induction Step: Assume that, for all players i, $(c_i, b_i^n) \in R_i^n(k+1)$ whenever (c_i, b_i^n) is consistent with up to k-fold belief in rationality. Now let (c_i, b_i^n) be consistent with up to $k+1$-fold belief in rationality. We need to show that $(c_i, b_i^n) \in R_i^n(k+2)$.

Since (c_i, b_i^n) is consistent with up to $k+1$-fold belief in rationality, there is a $b_i \in B_i$ that expresses up to $k+1$-fold belief in rationality such that b_i rationalizes c_i and induces b_i^n.

Hence, we know that

1. $u_i(c_i, b_i^{n+1}) \geq u_i(c_i', b_i^{n+1}), \forall c_i' \in C_i$ where b_i^{n+1} is induced by b_i.
2. b_i also expresses up to k-fold belief in rationality. So, by the induction assumption, $(c_i, b_i^n) \in R_i^n(k+1)$ where b_i^n is induced by b_i.
3. b_i assigns probability 1 to the set of opponents' choice-belief-hierarchy combinations (c_{-i}, b_{-i}), where, for

every $j \neq i$, b_j rationalizes c_j and expresses up to k-fold belief in rationality. So, by the induction assumption, for every such (c_j, b_j), we have that $(c_j, b_j^n) \in R_j^n(k+1)$, $j \neq i$ where b_j^n is induced by b_j and therefore $b_i^{n+1} \in \Delta(R_{-i}^n(k+1))$.

4. $b_i^n = \text{marg}_{C_{-i} \times B_{-i}^{n-1}} b_i^{n+1}$.

Combining (1)-(4), it follows that $(c_i, b_i^n) \in R_i^n(k+2)$, establishing the first direction.

$\Leftarrow$ For this direction, we show that, for any $(c_i, b_i^n) \in R_i^n(k+1)$, there is a belief hierarchy b_i exhibiting up to k-fold belief in rationality that induces b_i^n and rationalizes c_i. Again, we proceed by induction over $k \geq 0$.

Induction Start: Let $(c_i, b_i^n) \in R_i^n(1)$. Then there is a b_i^{n+1} that induces b_i^n and rationalizes c_i. So take any b_i such that b_i induces b_i^{n+1}. Then b_i rationalizes c_i.

Induction Step: Assume that, for every player i and any $(c_i, b_i^n) \in R_i^n(k+1)$, there is a belief hierarchy b_i inducing b_i^n, rationalizing c_i and exhibiting up to k-fold belief in rationality. We have to show that if $(c_i, b_i^n) \in R_i^n(k+2)$ then there is a belief hierarchy b_i that exhibits up to $k+1$-fold belief in rationality, induces b_i^n and rationalizes c_i.

So let $(c_i, b_i^n) \in R_i^n(k+2)$. Then there is an $n+1$th-order belief $b_i^{n+1} \in \Delta(R_{-i}^n(k+1))$ that rationalizes c_i and induces b_i^n. For every player $j \neq i$, let $\Theta_j^n \subseteq R_j^n(k+1)$ be the set of combinations of choices and nth-order beliefs in the support of b_i^{n+1}. By the induction assumption, for any $(c_j, b_j^n) \in \Theta_j^n$, we can pick a belief hierarchy $\hat{b}_j(c_j, b_j^n)$ that expresses up to k-fold belief in rationality, induces b_j^n and rationalizes c_j. Let $\theta_j(c_j, b_j^n) = (c_j, \hat{b}_j(c_j, b_j^n))$ be the resulting choice-belief-hierarchy combination. Given the mapping θ_j, for any measurable $E_j^n \subseteq \Theta_j^n$, let $\theta_j(E_j^n) = \{\theta_j(c_j, b_j^n) | (c_j, b_j^n) \in E_j^n\}$. Now let b_i be the belief hierarchy given by $b_i^{n+1}(\bigtimes_{j \neq i} E_j^n) = b_i(\bigtimes_{j \neq i} \theta_j(E_j^n))$ for every measurable $\bigtimes_{j \neq i} E_j^n \subseteq \bigtimes_{j \neq i} \Theta_j^n$. Since $\text{proj}_{C_j \times B_j^n} \theta_j(E_j^n) = E_j^n$ for all $E_j^n \subseteq \Theta_j^n$, this construction implies that b_i induces b_i^{n+1}. Moreover, b_i assigns probability 1 to combinations of choices and belief hierarchies $\theta_j(c_j, b_j^n)$ such that $\theta_j(c_j, b_j^n)$ expresses rationality and up to k-fold belief in rationality, and hence b_i expresses up to $k+1$-fold belief in rationality. And, lastly, as b_i induces b_i^{n+1} and b_i^{n+1} rationalizes c_i, b_i rationalizes c_i as well. This establishes the second direction.

Part 2:

To start, we show $\overline{R}_i^n \supseteq \text{proj}_{C_i \times B_i^n} R_i(\omega)$. In words, every (c_i, b_i^n) that is consistent with common belief in rationality survives transfinite elimination of choices and nth-order beliefs.

To see that this must be true, first note that any (c_i, b_i^n) that is consistent with common belief in rationality is, by definition, consistent with up to k-fold belief in rationality for any $k \geq 0$. Hence, by part 1 of the theorem $(c_i, b_i^n) \in \bigcap_{k \in 1,2,\ldots} R_i^n(k) = R_i^n(\omega)$.

Second, note that consistency of (c_i, b_i^n) with common belief in rationality implies that there is a choice-belief hierarchy combination (c_i, b_i) that expresses rationality and common belief in rationality and induces (c_i, b_i^n). Hence, in particular, the $n+1$th-order belief b_i^{n+1} that is induced by b_i assigns full probability to opponents' choice-nth-order-belief combinations (c_j, b_j^n) that are consistent with common belief in rationality. Invoking the argument from the previous paragraph, this implies that $(c_j, b_j^n) \in R_j(\omega)$ for all (c_j, b_j^n) in the support of b_i^{n+1} and hence $(c_i, b_i^n) \in R_i^n(\omega+1)$.

Clearly, we can repeat this argument for all (c_j, b_j^n) in the support of b_i^{n+1}. The same goes for all (c_k, b_k^n), $k \neq j$ in the support of b_j^{n+1} for each b_j^{n+1} such that (c_j, b_j^{n+1}) is in the support of the $n+2$th-order belief b_i^{n+2} induced by b_i. Continuing in this fashion and taking limits as needed, it now follows that $\overline{R}_i^n \supseteq \text{proj}_{C_i \times B_i^n} R_i(\omega)$.

For the reverse direction, note that, by definition of $\overline{R}_i^n$, we have iteratively eliminated choice-nth-order-belief combinations until the resulting reductions have converged.[22] Hence, we can write

$$\overline{R}_i^n = \Big\{ (c_i, b_i^n) \in \overline{R}_i^n \,|\, \exists b_i^{n+1} \in \Delta(\overline{R}_{-i}^n) \text{ with } \text{marg}_{C_{-i} \times B_{-i}^{n-1}} b_i^{n+1} = b_i^n \text{ s.t. } u_i(c_i, b_i^{n+1}) \geq u_i(c_i', b_i^{n+1}), \forall c_i' \in C_i \Big\}.$$

In words, $\overline{R}_i^n$ is a best-response set: For any $(c_i, b_i^n) \in \overline{R}_i^n$, we can find $b_i^{n+1} \in \Delta(\overline{R}_{-i}^n)$ such that b_i^{n+1} induces b_i^n and rationalizes c_i.

Moreover, since $b_i^{n+1} \in \Delta(\overline{R}_{-i}^n)$, for every (c_j, b_j^n) in the support of b_i^{n+1}, we can also find a $b_j^{n+1} \in \Delta(\overline{R}_{-j}^n)$ such that b_j^{n+1} induces b_j^n and rationalizes c_j.

Continuing in this fashion, we construct an infinite sequence $b_i^{n+1}, b_i^{n+2}, \ldots$ of $n+k$th-order beliefs where $b_i^{n+1} \in \Delta(\overline{R}_{-i}^n)$ induces b_i^n and rationalizes c_i, b_i^{n+2} induces b_i^{n+1}, rationalizes c_i, and assigns full probability to $(c_j', b_j^{n+1}), j \neq i$ such that c_j' is rational given b_j^{n+1} and $b_j^{n+1} \in \Delta(\overline{R}_{-j}^n)$, and so on.

Taking $b_i^{n+1}, b_i^{n+2}, \ldots$ and adding c_i as well as the marginal distributions $b_i^1, \ldots, b_i^n$ induced by b_i^{n+1}, we arrive at a choice-belief-hierarchy tuple (c_i, b_i) that expresses common belief in rationality and induces (c_i, b_i^n). Since (c_i, b_i^n) was arbitrary, it follows that $\text{proj}_{C_i \times B_i^n} R_i(\omega) \supseteq \overline{R}_i^n$.

Part 3:

$R_i^n(\omega) \supseteq \text{proj}_{C_i \times B_i^n} R_i(\omega)$ was already proved in part 2.

To show that the reverse direction also applies in belief-continuous games, assume that $(c_i, b_i^n) \in R_i^n(\omega)$. Then (c_i, b_i^n) is consistent with up to k-fold belief in rationality for any $k \geq 0$ (where $k = 0$ means rational choice). Let $B_i[b_i^n]$ be the set of belief hierarchies that induce b_i^n. Note that, since $B_i[b_i^n]$ is closed and B_i is a compact Hausdorff space,[41] $B_i[b_i^n]$ is compact.

Next, for any $k \geq 0$, let $B_i[k, c_i]$ be the set of belief hierarchies that rationalize c_i and express up to k-fold belief in rationality.

To prove the lemma, we show that $B_i[k, c_i]$ is a compact set for every $k \geq 0$. Since the sequence $B_i[0, c_i] \cap B_i[b_i^n]$, $B_i[1, c_i] \cap B_i[b_i^n]$, $\ldots$ of belief hierarchies that rationalize c_i under up to k-fold belief in rationality and induce b_i^n is then a decreasing sequence of nested non-empty compact sets, Cantor's intersection theorem implies that $\bigcap_{k \in \{0,1,2,\ldots\}} \big(B_i[k, c_i] \cap B_i[b_i^n]\big)$ is non-empty such that (c_i, b_i^n) is indeed consistent with common belief in rationality.

We now show, by induction over $k \geq 0$, that every $B_j[k, c_j]$ is compact and metrizable for every player j, every $c_j \in C_j$ and every $k \geq 0$:

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

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

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

41 Compact Hausdorffness follows from the finiteness of all C_i, $i \in I$ and Tychonoff's theorem.

By the induction assumption, $\bigtimes_{\ell \neq j}\{(c_\ell, b_\ell) | b_\ell \in B_\ell[k, c_\ell]\}$ 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 $\Delta(\bigtimes_{\ell \neq j}\{(c_\ell, b_\ell) | b_\ell \in B_\ell[k, c_\ell]\})$. It follows that $B_j[k+1, c_j]$ is compact and metrizable, completing the induction and hence the proof. □

Proof of Observation 6.1*(Sketch)*.
Possibility: If rational choices exist for every belief hierarchy b_i, use the methods from the constructive proof of Theorem 4.3 to construct probability-one belief hierarchies expressing up to k-fold belief in rationality for any finite $k \geq 1$. This yields Part 1. Next, assuming that all choice sets C_i are compact Hausdorff, and that all u_i are belief-continuous, Part 2 follows as a corollary of Part 1 and (Part 2 of) the procedural characterization proven below.
Procedural Characterization: For Part 1, first note that bounded and measurable utilities u_i and separable choice sets C_i for all $i \in I$ are enough to guarantee that $R_i(k)$, $k \geq 1$ and (hence) $R_i(\omega)$ are measurable subsets of $S_i \times B_i$. The arguments from Theorem 5.5, Part 1 and Part 2 will then essentially go through. The main complication is that, given a purely measurable space of choices and belief hierarchies as in Heifetz and Samet (1998), order-$n+1$ belief-finiteness of all u_i does not guarantee that $R_i^n(k)$ is measurable for every $k \geq 1$, which necessitates a more involved definition of *elimination of choices and nth-order beliefs*. See Jagau (2021) for details and proofs.

Given Part 1, assuming that all choice sets C_i are compact Hausdorff and that Γ is belief-continuous, Theorem 5.5, Part 3 straightforwardly generalizes,[42] thus establishing Part 2. □

Appendix B. Transfinite elimination of choices and *n*th-order beliefs

Here we present an example of a three-player belief-discontinuous psychological game that is belief-finite of order two. As we will show, elimination of choices and first-order beliefs (Procedure 5.4) takes $\omega + 1$ steps to determine the tuples of choices and first-order beliefs that are consistent with common belief in rationality here. The structure of our example is similar to Example 2 in Dufwenberg and Stegeman (2002), which uses a traditional game with uncountably many choices for two of out of three players. By contrast, the game in our example has just two choices for each of the three players.

Example B.1 (*Transfinite Elimination of Choices and First-Order Beliefs*)**. Two-Tiered Battle of the Sexes with an Outside Wager:**
We consider a $2 \times 2 \times 2$-psychological game between you, Alice, and Bob. Formally, let $I = \{y, a, b\}$, and let the choice sets be $C_y = \{B, F\}$, $C_a = \{B, F\}$, $C_b = \{CBR, \neg CBR\}$. Your utility function is given by

$$u_y(B, b_y) = 2(b_y^1(B) + e_y^2(B, B) + e_y^2(F, B)) \text{ and}$$
$$u_y(F, b_y) = (b_y^1(F) + e_y^2(B, F) + e_y^2(F, F))$$

where we define $e_y^2(c_a, c_y) := \int_{\{c_a\} \times B_a} b_a^1(c_y)\, db_y$ for every $(c_y, c_a) \in C_y \times C_a$. This expression captures the *expected probability* which you believe Alice assigns to your choice c_y conditional on choosing c_a herself. This amounts to a summary statistic of your second-order belief b_y^2 which we may call the *second-order expectation*.[20] Analogously, Alice's utility function is given by

$$u_a(B, b_a) = (b_a^1(B) + e_a^2(B, B) + e_a^2(F, B)) \text{ and } u_a(F, b_a)$$
$$= 2(b_a^1(F) + e_a^2(B, F) + e_a^2(F, F)).$$

Since the utility functions for you and Alice are *additively separable* in first-order beliefs and second-order expectations, we can represent each of them as a sum of two payoff matrices:

The total utilities for you and Alice are the sum of the two utility components. For instance, your utility from choosing B if your first-order belief b_y^1 is B and your second-order expectation e_y^2 is $\frac{1}{2}(B, B) + \frac{1}{2}(B, F)$ is equal to $2 + \frac{1}{2}(2 + 0) = 3$.

One way to think about the game between you and Alice is as a two-tiered version of the Battle of the Sexes. That is, both you and Alice want to coordinate on the same choice and, independently, you both want to live up to your co-player's expectations regarding your behavior ($\hat{=}$ second-order expectations). In addition, conditional on behavioral or expectational coordination, you have a preference for choice B whereas Alice has an analogous preference for choice F.

Now besides you and Alice there is a third player – *Bob* – that observes the game between you and Alice.[43] To concisely write down Bob's utility function, let $b_{b,a}^2 = \text{marg}_{C_a \times B_a^1} b_b^2$. In words, $b_{b,a}^2$ is Bob's second-order belief regarding only Alice's choice and first-order belief. Also, define the set $\overline{R}_a^1 := (\{F\} \times B_a^1) \cup (\{B\} \times \{b_a^1 \in B_a^1 | b_a^1(B) \geq \frac{1}{\sqrt{3}}\})$. Bob's utility function is then described by the following matrix:

As will become clear below and as suggested by our naming of Bob's choices, Bob can be thought of as taking a bet on whether Alice's second-order belief is consistent with common belief in rationality.

Note that Bob's utility function is not belief-continuous. For example, perturbing Bob's second-order beliefs about Alice's choice and first-order belief slightly around a degenerate belief that assigns full probability to $(c_a, b_a^1) = (B, \frac{1}{\sqrt{3}}B + \frac{\sqrt{3}-1}{\sqrt{3}}F)$ can make $u_b(CBR, b_b^2)$ jump discontinuously from 1 to 0.

We will now show that iterated elimination of choices and first-order beliefs eliminates Bob's choice $\neg CBR$, but only at the $\omega + 1$th step of elimination.

To this end, we first argue that the sets $R_a^1(k)$ of Alice's choice-first-order-belief tuples consistent with rationality and up to k-fold belief in rationality continuously shrink towards the set $\overline{R}_a^1$ while iterating over all finite k. Start by considering Alice's utility maximization problem. Alice prefers her choice B whenever

$$\begin{aligned} u_a(B, b_a^2) - u_a(F, b_a^2) &= b_a^1(B) - 2(1 - b_a^1(B)) + (e_a^2(B, B) \\ &\quad + e_a^2(F, B)) - 2(e_a^2(B, F) + e_a^2(F, F)) \qquad (1) \\ &= 3(b_a^1(B) + e_a^2(B, B) + e_a^2(F, B)) - 4 \geq 0, \end{aligned}$$

where we used $b_a^1(F) = 1 - b_a^1(B)$, $e_a^2(B, F) = b_a^1(B) - e_a^2(B, B)$, and $e_a^2(F, F) = (1 - b_a^1(B)) - e_a^2(B, F)$. Now if Alice's second-order expectation is such that $e_a^2(B, B) + e_a^2(F, B) = 0$, the inequality reduces to $3b_a^1(B) - 4 \geq 0$, which is impossible given that $b_a^1(B) \in [0, 1]$. This shows that the tuple (F, b_a^1) can be rationalized for any of Alice's first-order beliefs $b_1^a \in B_1^a$.

On the other hand, for any fixed first-order belief $b_a^1 \in B_a^1$, we can observe that the difference $u_a(B, b_a^2) - u_a(F, b_a^2)$ is maximal for $e_a^2(B, B) + e_a^2(F, B) = 1$.

Hence, the lowest probability that Alice can ever assign to your choice B while rationally choosing B follows from

$$3b_a^1(B) - 1 \geq 0 \Leftrightarrow b_a^1(B) \geq \frac{1}{3}.$$

[42] That all B_i and, a fortiori, all spaces of probability-one belief hierarchies are compact Hausdorff is a straightforward implication of compact Hausdorffness of C_i for every player i and Tychonoff's Theorem.

[43] For ease of notation, we will refrain from modeling beliefs that you and Alice might have regarding Bob's behavior and beliefs, beliefs that you, Alice, and Bob might have regarding these beliefs, and so on. Since Bob is merely a bystander with respect to the game between you and Alice, this simplification does not affect our results.

Summing up, we have $R_a^1(1) = (\{F\} \times B_a^1) \cup (\{B\} \times \{b_a^1 \in B_a^1 | b_a^1(B) \geq \frac{1}{3}\})$, and, with the symmetry of the game between you and Alice, we find $R_y^1(1) = (\{B\} \times B_y^1) \cup (\{F\} \times \{b_y^1 \in B_y^1 | b_y^1(F) \geq \frac{1}{3}\})$.

Next, if Alice believes in your rationality, we must have $b_a^2 \in \Delta(R_y^1(1))$. Concretely, this means that Alice cannot believe you to choose F while assigning probability less than $\frac{1}{3}$ to Alice choosing F as well. In terms of Alice's second-order expectation and first-order belief this leads to the constraint $e_a^2(F,F) \geq \frac{1}{3} b_a^1(F) \Leftrightarrow \frac{2}{3} b_a^1(F) \geq e_a^2(F,B)$. Going back to Alice's utility maximization problem from Eq. (1), we see that the constraint is irrelevant for rationalizing Alice's choice F (which is most attractive for $e_a^2(B,F) = 0$). On the other hand, for any first-order belief b_a^1 for Alice, the maximal utility difference in favor of choice B is now attained precisely where

$e_a^2 = (b_a^1(B)(B,B) + \frac{2b_a^1(F)}{3}(F,B) + \frac{b_a^1(F)}{3}(F,F))$. Hence, the lowest probability that Alice can ever assign to your choice B while rationally choosing B under belief in rationality follows from

$$3\left(2b_a^1(B) + \frac{2}{3}(1 - b_a^1(B))\right) - 4 = 4b_a^1(B) - 2 \geq 0 \Leftrightarrow b_a^1(B) \geq \frac{1}{2}.$$

Hence, we have $R_a^1(2) = (\{F\} \times B_a^1) \cup (\{B\} \times \{b_a^1 \in B_a^1 | b_a^1(B) \geq \frac{1}{2}\})$ and (again by symmetry) $R_y^1(1) = (\{B\} \times B_y^1) \cup (\{F\} \times \{b_y^1 \in B_y^1 | b_y^1(F) \geq \frac{1}{2}\})$. Continuing in this fashion, it is now clear that the sets of first-order beliefs supporting Alice's choice F and your choice B will be the same for all levels of up to k-fold belief in rationality, whereas the sets of first-order beliefs supporting Alice's choice B and your choice F will get smaller and smaller as we iterate k over the natural numbers.

To describe the law of motion of $R_a^1(k)$, $k \in \omega$, for any finite $k \geq 0$, let $b_y^1(F,k)$ denote the minimum probability you must assign to Alice's choice F while rationally choosing F under up to k-fold belief in rationality, and analogously define $b_a^1(B,k)$. Starting from Eq. (1), for any first-order belief b_a^1 for Alice, the maximal utility difference in favor of choice B under up to k-fold belief in rationality is attained precisely where $e_a^2 = (b_a^1(B)(B,B) + (1 - b_y^1(F,k-1)) b_a^1(F)(F,B) + b_y^1(F,k-1) b_a^1(F)(F,F))$, so that $b_a^1(B,k)$ follows from

$$\begin{aligned} &3\left(2b_a^1(B) + (1 - b_y^1(F,k-1))(1 - b_a^1(B))\right) - 4 \\ &\quad = (1 + b_y^1(F,k-1)) b_a^1(B) - 3b_y^1(F,k-1) - 1 \geq 0 \Leftrightarrow \\ &\quad b_a^1(B) \geq 1 - \frac{2}{3(1 + b_y^1(F,k-1))} = b_a^1(B,k) \end{aligned}$$

Using again the symmetry of the game between you and Alice, we can identify $b_y^1(F,k-1) = b_a^1(B,k-1)$, leading to the first-order difference equation

$$b_a^1(B,k) = 1 - \frac{2}{3(1 + b_a^1(B,k-1))}$$

with initial condition $b_a^1(B,0) = 0$. It is straightforward to show that $b_a^1(B,k)$ increases in k at a decreasing rate, and that $0 < \sup_{k \in \omega} b_a^1(B,k) < 1$.

Letting $b^* := \sup_{k \in \omega} b_a^1(B,k)$, it then follows that

$$b^* = 1 - \frac{2}{3(1 + b^*)} \Leftrightarrow b^* = \frac{1}{\sqrt{3}}$$

Thus, the tuples of choices and first-order beliefs that are consistent with common belief in rationality for Alice are given by

$$\begin{aligned} R_a^1(\omega) &= \bigcap_{k \in \omega} R_a^1(k) = (\{F\} \times B_a^1) \cup \\ &\times \left(\{B\} \times \left\{b_a^1 \in B_a^1 \middle|\ b_a^1(B) \geq \frac{1}{\sqrt{3}}\right\}\right) = \overline{R}_a^1. \end{aligned}$$

It is now easy to see that Bob's choice $\neg CBR$ will be eliminated at step $\omega + 1$ of elimination of choices and first-order beliefs: Since $R_a^1(k) \supset \overline{R}_a^1$ for any finite k, there is a second-order belief for Bob that makes his choice $\neg CBR$ consistent with up to k-fold belief in rationality for any finite k, and hence $\neg CBR \in \text{proj}_{C_b}(R_b^1(\omega))$. However, as soon as we require $b_{b,a}^2 \in \Delta(R_a^1(\omega)) = \Delta(\overline{R}_a^1)$, choice CBR is strictly better than $\neg CBR$ for Bob. So we must indeed have $\text{proj}_{C_b}(R_b^1(\omega + 1)) = \{CBR\}$, as claimed.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmateco.2022.102635.

References

Asheim, G.B., Perea, A., 2005. Sequential and quasi-perfect rationalizability in extensive games. Games Econ. Behav. 53 (1), 15–42.

Attanasi, G., Battigalli, P., Manzoni, E., 2016. Incomplete-information models of guilt aversion in the trust game. Manage. Sci. 62 (3), 648–667.

Attanasi, G., Battigalli, P., Manzoni, E., Nagel, R., 2019. Belief-dependent preferences and reputation: Experimental analysis of a repeated trust game. J. Econ. Behav. Organ. 167, 341–360.

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

Battigalli, P., 1997. On rationalizability in extensive games. J. Econ. Theory 74 (1), 40–61.

Battigalli, P., Corrao, R., Sanna, F., 2020. Epistemic game theory without types structures: An application to psychological games. Games Econ. Behav. 120, 28–57.

Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Amer. Econ. Rev. 97 (2), 170–176.

Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. J. Econ. Theory 144 (1), 1–35.

Battigalli, P., Dufwenberg, M., Smith, A., 2019. Frustration, aggression, and anger in leader-follower games. Games Econ. Behav. 117, 15–39.

Battigalli, P., Siniscalchi, M., 2002. Strong belief and forward induction reasoning. J. Econ. Theory 106 (2), 356–391.

Ben-Porath, E., 1997. Rationality, Nash equilibrium and backwards induction in perfect-information games. Rev. Econ. Stud. 64 (1), 23–24.

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

Bjorndahl, A., Halpern, J.Y., Pass, R., 2013. Language-based games. In: Proceedings of the Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge. pp. 39–48, extended working paper available at http://www.cs.cornell.edu/~rafael/papers/lbg.pdf.

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

Brandenburger, A., Dekel, E., 1993. Hierarchies of beliefs and common knowledge. J. Econ. Theory 59 (1), 189–198.

Caplin, A., Leahy, J., 2004. The supply of information by a concerned expert. Econom. J. 114 (497), 487–505.

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74 (6), 1579–1601.

Chen, Y.-C., Long, N.V., Luo, X., 2007. Iterated strict dominance in general games. Games Econ. Behav. 61 (2), 299–315.

Dekel, E., Fudenberg, D., Levine, D.K., 1999. Payoff information and self-confirming equilibrium. J. Econ. Theory 89 (2), 165–185.

Dekel, E., Fudenberg, D., Levine, D.K., 2002. Subjective uncertainty over behavior strategies: A correction. J. Econ. Theory 104 (2), 473–478.

Dufwenberg, M., 2002. Marital investments, time consistency and emotions. J. Econ. Behav. Organ. 48 (1), 57–69.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47 (2), 268–298.

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

Dufwenberg, Jr., M., Dufwenberg, M., 2018. Lies in disguise: A theoretical analysis of cheating. J. Econ. Theory 175, 248–264.

Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1 (1), 60–79.

Heifetz, A., Mongin, P., 2001. Probability logic for type spaces. Games Econ. Behav. 35 (1–2), 31–53.

Heifetz, A., Samet, D., 1998. Topology-free typology of beliefs. J. Econ. Theory 82 (2), 324–341.

Heifetz, A., Samet, D., 1999. Coherent beliefs are not always types. J. Math. Econ. 32, 475–488.

Huck, S., Kübler, D., 2000. Social pressure, uncertainty, and cooperation. Econ. Gov. 1, 199–212.

Jagau, S., 2021. The Fundamental Theorem of Epistemic Game Theory: The Infinite Case. Epicenter Working Paper No. 25, Link: http://www.epicenter.name/wp-content/uploads/2021/12/EPICENTER_WP_25.pdf.

Jagau, S., Perea, A., 2017. Common Belief in Rationality in Psychological Games. Epicenter Working Paper No. 10, Link: http://www.epicenter.name/wp-content/uploads/2017/09/Psych-Games-CBR-WP.pdf.

Jagau, S., Perea, A., 2021. Linear Psychological Games. Working paper.

Khalmetski, K., Ockenfels, A., Werner, P., 2015. Surprising gifts: Theory and laboratory evidence. J. Econ. Theory 159, 163–208.

Kolpin, V., 1992. Equilibrium refinement in psychological games. Games Econ. Behav. 4 (2), 218–231.

Kreps, D.M., Wilson, R., 1982. Sequential equilibria. Econometrica 50 (4), 863–894.

Li, J., 2008. The power of conventions: A theory of social preferences. J. Econ. Behav. Organ. 65 (3), 489–505.

Lipman, B.L., 1994. A note on the implications of common knowledge of rationality. Games Econ. Behav. 6 (1), 114–129.

Meier, M., 2012. An infinitary probability logic for type spaces. Isr. J. Math. 192, 1–58.

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

Mourmans, N., 2019. Reasoning in Psychological Games: When is Iterated Elimination of Choices Enough? Epicenter Working Paper No. 20, Link: https://www.epicenter.name/wp-content/uploads/2019/06/When-is-iterated-elimination-of-choices-enough-Epicenter.pdf.

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

Perea, A., 2014. Belief in the opponents future rationality. Games Econ. Behav. 83, 231–254.

Rabin, M., 1993. Incorporating fairness into game theory and economics. Amer. Econ. Rev. 83 (5), 1281–1302.

Sebald, A., 2010. Attribution and reciprocity. Games Econ. Behav. 68 (1), 339–352.

Spohn, W., 1982. How to make sense of game theory? In: Stegmüller, W., Balzer, W., Spohn, W. (Eds.), Philosophy of Economics. Springer, Berlin/Heidelberg, pp. 239–270.

Tan, T.C.-C., da Costa Werlang, S.R., 1988. The Bayesian foundations of solution concepts of games. J. Econ. Theory 45 (2), 370–391.

Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211 (4481), 453–458.

Wakker, P.P., 2010. Prospect Theory: For Risk and Ambiguity. Cambridge University Press, Cambridge.