Common Belief in Rationality in Psychological Games

Stephan Jagau[A B C] and Andrés Perea[B D]

EPICENTER Working Paper No. 10 (2017)

EPICENTER
Research Center for Epistemic Game Theory

Belief-dependent motivations and emotional mechanisms such as surprise, anxiety, anger, guilt, and intention-based reciprocity pervade real-life human interaction. At the same time, traditional game theory has experienced huge difficulties trying to capture them adequately. Psychological game theory, initially introduced by Geanakoplos et al. (1989), has proven to be a useful modeling framework for these and many more psychological phenomena. In this paper, we use the epistemic approach to psychological games to systematically study common belief in rationality, also known as correlated rationalizability. We show that common belief in rationality is possible in any game that preserves rationality at infinity, a mild requirement that is considerably weaker than the previously known continuity conditions from Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009). Also, we provide an example showing that common belief in rationality might be impossible in games where rationality is not preserved at infinity. We then develop an iterative procedure that, for a given psychological game, determines all rationalizable choices. In addition, we explore classes of psychological games that allow for a simplified procedure.

JEL classification: C72, D03, D83

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

---

[A]CREED, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
[B]EpiCenter, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands
[C]Email: S.D.Jagau@uva.nl. Web: https://sites.google.com/view/stephanjagau
[D]Email: a.perea@maastrichtuniversity.nl. Web: http://www.epicenter.name/Perea/

I Introduction and Related Literature

Traditional game theory rests on the assumption that decision-makers exclusively care about the outcomes that materialize as a result of their choices and the choices of their opponents. However, in many real-life interactions, we can see ourselves caring not only about outcomes, but also about our anticipated emotional reactions and the beliefs, opinions, and emotional reactions of others. In short: Intentions matter for how we choose to act and outcome-based preferences as used in traditional game theory give us a hard time trying to capture this aspect of human behavior. *Psychological game theory*, pioneered by Geanakoplos et al. (1989) and more recently extended to sequential interaction by 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 its introduction, the psychological games framework has proven to be a useful tool for many applications in behavioral and experimental economics. It has been used to model belief-dependent motivations so diverse as intention-based reciprocity (Rabin 1993, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006, Sebald 2010), guilt (Huang and Wu 1994, Dufwenberg 2002, Charness and Dufwenberg 2006, Battigalli and Dufwenberg 2007, Attanasi et al. 2016, Attanasi et al. 2017), social pressure and conformity (Huck and Kübler 2000, Li 2008), anxiety (Caplin and Leahy 2004), lying aversion (Battigalli et al. 2013, Dufwenberg and Dufwenberg 2016), surprise (Khalmetski et al. 2015), anger (Battigalli et al. 2015), and esteem (Akerlof 2017).

At the same time, theoretical work on psychological games has largely remained explorative. Early results by Geanakoplos et al. (1989) and Kolpin (1992) are concerned with generalizing Nash equilibrium and various refinements to psychological games and provide sufficient conditions for existence of these equilibria. Battigalli and Dufwenberg (2009) formally extend the psychological games framework to sequential interaction and provide a definition of common strong belief in rationality (Battigalli and Siniscalchi 2002, characterizing extensive-form rationalizability, Pearce 1984) and sequential equilibrium (Kreps and Wilson 1982) for dynamic psychological games.

Notwithstanding these first steps towards a systematic theoretical treatment of psychological games, many fundamental questions remain unaddressed. This is true even for the most basic mode of reasoning in games, common belief in rationality (Brandenburger and Dekel 1987, Tan and Werlang 1988, characterizing correlated rationalizability).

Applications of rationalizability in the applied analysis of specific psychological games have previously been presented in, among others, Battigalli and Dufwenberg (2009), Battigalli et al. (2013), and Attanasi et al. (2016). Also, Battigalli and Dufwenberg (2009) define common strong belief in rationality (reducing to common belief in rationality in static games) for arbitrary dynamic psychological games and provide an existence condition akin to Geanakoplos et al.'s (1989) continuity condition.

In this paper, we extend and systematize what was previously known by providing an extensive treatment of common belief in rationality in arbitrary static psychological games. In particular, we provide an algorithmic characterization of rationalizability for all static psychological games which has so far been absent from the published literature. Also, we present a novel existence condition for common belief in rationality in static psychological games that considerably weakens the previously known continuity condition.

We firstly examine the possibility of common belief in rationality in psychological games.

We show that common belief in rationality is possible in any psychological game that *preserves rationality at infinity*. That is, if a choice is irrational for a given belief hierarchy, we can point to a finite order of beliefs to expose the irrationality of that choice-belief-hierarchy combination. This result is similar to the condition *CR* for the existence of rationalizable strategies in *language-based games* presented in a recent working paper by Bjorndahl et al. (2016) on language-based games. This class of games includes some, but not all, psychological games as usually defined. Specifically, the psychological games that can be mapped into language-based games would only allow players to entertain deterministic belief hierarchies and linear combinations of such deterministic belief hierarchies. By contrast, we will allow players to entertain all possible probabilistic belief hierarchies as is common in the psychological-games literature. Hence, our condition of preservation of rationality at infinity may be viewed as an extension of Bjorndahl et al.'s (2016) CR-condition to a broader class of psychological games.

Special cases of games that preserve rationality at infinity are *belief-finite psychological games* where players' utilities depend on finitely many levels of higher-order beliefs and psychological games where players' utilities are continuous functions of belief hierarchies in the sense of the weak topology (cf. Geanakoplos et al. 1989, Battigalli and Dufwenberg 2009). In addition to the existence condition, we also provide an example showing that common belief in rationality might be impossible whenever a game does not preserve rationality at infinity.

Secondly, we develop an iterative elimination procedure over choices and belief hierarchies that characterizes common belief in rationality for all psychological games. Our procedure generalizes iterated elimination of strictly dominated choices as used in traditional games in an intuitive way. However, 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 the algorithm for general psychological games. A substantial part of this paper and a companion paper (Jagau and Perea 2017) are therefore devoted to studying classes of games that allow for a simplified procedure. In particular, we will provide conditions under which the applicable algorithm *does* share the attractive linearity and finiteness properties of iterated elimination of strictly dominated choices.

For *belief-finite games* where player's utilities depend on at most nth-order beliefs, we find that iterative elimination of choices and $n-1$th-order beliefs characterizes common belief in rationality.

Next to our paper, an unpublished master thesis by Sanna (2016) provides an algorithmic characterization of common belief in rationality for static psychological games where utilities depend on finitely many levels.[1] While the procedure is very similar to ours, there are two crucial differences. Firstly, Sanna (2016) restricts to games satisfying the continuity condition originally introduced in Geanakoplos et al. (1989) while we show that continuous utility functions are not necessary either to prove the characterization result or to establish the possibility of common belief in rationality in a belief-finite game, even though it is necessary to use a more complex algorithm in the discontinuous case. Secondly, Sanna (2016) allows for possibly incoherent beliefs while our definition of a static psychological game rests on the more standard assumption that players' beliefs satisfy coherency and common belief in coherency.

A special case of belief-finite games that has been studied intensively but informally in applications of psychological game theory are *expectation-based games* in which players only care about expected values of finite levels of higher-order beliefs (cf. e.g. Rabin 1993, Falk and Fischbacher 2006, Battigalli and Dufwenberg 2007, Dufwenberg and Dufwenberg 2016). In our companion paper Jagau and Perea (2017), we provide a first formal definition of this class of psychological games. Also we introduce the subclass of *expectation-based games* in which utilities *linearly* depend on expectations. We show that these games are *additive*, that is, they admit a natural generalization of the expected utility representation within the realm of psychological games. This comes hand in hand with a matrix representation of utility and an LP-implementable procedure characterizing common belief in rationality, the latter under the additional condition that utility depends on finitely many level of beliefs. Existing models that fall into the class of additive game include simple guilt (Battigalli and Dufwenberg 2007) and simple surprise (Khalmetski et al. 2015).

Another special class of belief-finite games that we study in the present paper are unilateral games where one player cares about second-order beliefs and all others care about first-order beliefs only. For this class, which also surfaces in numerous applications of psychological game theory (cf. e.g. Huang and Wu 1994, Dufwenberg 2002, Charness and Dufwenberg 2006, Battigalli and Dufwenberg 2007, 2009), we show that common belief in rationality is characterized by a finite procedure.

The remainder of this paper is structured as follows: Section II introduces the psychological-games framework. Section III extends the definition of common belief in rationality to psychological games. Section IV provides sufficient conditions for common belief in rationality to be possible in a given psychological game. Section V develops the iterative belief-elimination procedure that characterizes common belief in rationality in psychological games. The remaining sections study classes of games in which common belief in rationality can be characterized by a simplified algorithm: In section VI, we introduce the algorithm *iterated elimination of choices and nth-order*

[1] In a second part of that thesis, the algorithm is extended to common strong belief in rationality for dynamic psychological games.

beliefs for *belief-finite psychological games* in which players only care about higher-order beliefs up to some finite order. In section VII, we study the class of unilateral games, in which exactly one player cares about second-order beliefs and all other players have standard preferences. For these games, we present the finite algorithm *iterated elimination of choices and opponents' 1st-order beliefs*. Lastly, section VIII compares our approach to modeling psychological games with other models used in the literature, summarizes our findings, and concludes with a systematic classification of psychological games.

II Psychological Games

In this section, we start by providing a formal definition of static psychological games. The definition is entirely self-contained and concludes with an explicit comparison between traditional games and psychological games.

In a traditional game, players' utilities depend only on their choices and their first-order beliefs about the opponents' choices and, moreover, they depend *linearly* on the first-order beliefs. By contrast, utilities in general psychological games might depend *non-linearly* on the full *belief hierarchy* of players. Before giving a formal definition of psychological games, we therefore define what is meant by a belief hierarchy of a given player: Each belief hierarchy b_i is a chain of probability distributions that capture i's belief about his opponents' choices, his beliefs about his opponents' beliefs about their opponents' choices and so on and so forth. Each level $n \geq 1$ of this chain is represented by an nth-order belief b_i^n. Brandenburger and Dekel (1993) show how the sets of nth-order beliefs can be recursively constructed. Following their approach, for any polish space S, let $\Delta(S)$ denote the set of probability measures on the Borel-field over S and endow $\Delta(S)$ with the weak topology. In our case, the relevant space of uncertainty for player i is the set of opponents' choices $\times_{j \neq i} C_j = C_{-i}$.

We start by defining the sets

$$
\begin{aligned}
X_i^1 &= C_{-i} \\
X_i^2 &= X_i^1 \times \bigtimes_{j \neq i} \Delta(X_j^1) \\
&\vdots \\
X_i^n &= X_i^{n-1} \times \bigtimes_{j \neq i} \Delta(X_j^{n-1}) \\
&\vdots
\end{aligned}
$$

Let $\tilde{B}_i(0) = \times_{n=1}^{\infty} \Delta(X_i^n)$ be the set of all belief hierarchies for player i. For every belief hierarchy $b_i = (b_i^1, b_i^2, ...)$, the probability distribution $b_i^n \in \Delta(X_i^n)$ is called the nth-order belief of player i. If we want beliefs of players not to be self-contradictory, b_i cannot be just any element of $\tilde{B}_i(0)$.

Instead, it should satisfy *coherency*: For each b_i^n, $n \geq 2$, by marginalizing i's beliefs w.r.t. X_i^{n-1}, we should receive b_i^{n-1}.

Definition II.1. *(Coherency)*
A belief hierarchy $b_i = (b_i^1, b_i^2, ...)$ *is* **coherent** *if for every* $n \geq 2$*, it satisfies*

$$\text{marg}_{X_i^{n-1}} b_i^n = b_i^{n-1}.$$

Let $\tilde{B}_i(1) \subseteq \tilde{B}_i(0)$ *be the set of player* i*'s coherent beliefs.*

On top of this, no player should entertain beliefs that questions opponents' coherency at any level, i.e. we should restrict to i's belief hierarchies expressing *common belief in coherency.*

Using Brandenburger and Dekel's (1993) Proposition 1, we know that there is a homeomorphism $f_i : \tilde{B}_i(1) \to \Delta(C_{-i} \times \tilde{B}_{-i}(0))$. This allows us to iteratively construct B_i via

$$\tilde{B}_i(k) := \{b_i \in \tilde{B}_i(k-1) |\ f_i(b_i)(C_{-i} \times \tilde{B}_{-i}(k-1)) = 1\},\ k \geq 2$$

and $B_i = \bigcap_{k \geq 0} \tilde{B}_i(k)$. Then B_i contains i's belief hierarchies that express coherency and common belief in coherency.

For all $n \geq 1$, the set B_i^n of nth-order beliefs for player i that are consistent with coherency and common belief in coherency is given by $B_i^n = \text{proj}_{\Delta(X_i^n)} B_i$.

Before going on, note that Brandenburger and Dekel's (1993) Proposition 2 implies that every $b_i \in B_i$ is homeomorphic to a probability distribution in $\Delta(C_{-i} \times B_{-i})$. Therefore, whenever convenient, we will identify $b_i \in B_i$ with its corresponding probability distribution in $\Delta(C_{-i} \times B_{-i})$. Similarly, it is well known that also each $b_i^n \in B_i^n$ is homeomorphic to a probability distribution in $\Delta(C_{-i} \times B_{-i}^{n-1})$, allowing us to also identify $b_i^n \in B_i^n$ with its corresponding probability distribution in $\Delta(C_{-i} \times B_{-i}^{n-1})$ whenever that is useful.
We are now ready to define static psychological games:

Definition II.2. *(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 utility function of the form*

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

The way of modeling psychological games used here is slightly different from what has been done in the previous literature. In section VIII, we therefore examine how our definition of static psychological games relates to the two best-known previous definitions, namely the ones in Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009). As it turns out, our definition is entirely equivalent to these alternative definitions.

Before proceeding, it is useful to clarify how psychological games generalize traditional games. We need to impose two restrictions on a psychological game to receive a traditional static game.

First, we must have $u_i(c_i, b_i) = u_i(c_i, b_i')$ whenever $b_i^1 = b_i^{1\prime}$. In words, utility depends only on players' first-order beliefs while in general psychological games, it may depend on beliefs of arbitrary levels. We can then write utility as a function $u_i : C_i \times \Delta(C_{-i}) \to \mathbb{R}$.

Second, it must be the case that utility is linear in first-order beliefs or, equivalently, expected utility must hold. Formally, there must exist a function $v_i : C_i \times C_{-i} \to \mathbb{R}$ (Bernoulli utility) such that $u_i(c_i, b_i) = \sum_{c_{-i} \in C_{-i}} b_i^1(c_{-i}) v_i(c_i, c_{-i})$.

By contrast, utilities in general psychological games might depend non-linearly on beliefs of arbitrary order.

III Common Belief in Rationality

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

Definition III.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 III.1, we can define belief in the opponents' rationality. For this purpose, define the set $(C_i \times B_i)^{rat} := \{(c_i, b_i) \in C_i \times B_i |\ c_i \text{ is rational given } b_i\}$ of choice-belief combinations (c_i, b_i) such that the choice c_i is rational given belief hierarchy b_i.

Definition III.2. *(Belief in the Opponents' Rationality)*
Consider a belief hierarchy $b_i \in B_i$ for player i. Belief hierarchy b_i is said to express **belief in the opponents' rationality** *if $b_i \in \Delta(\times_{j \neq i} (C_j \times B_j)^{rat})$. In words, b_i assigns full probability to the set of opponents' choice-belief combinations where the choice is rational given the belief hierarchy.*

Going on from here, we define higher-order belief in the opponents' rationality and common belief in rationality:

Definition III.3. *(Up to k-Fold and Common Belief in Rationality)*
Recursively define

$$B_i(1) = \{b_i \in B_i |\ b_i \in \Delta(\bigtimes_{j \neq i} (C_j \times B_j)^{rat})\}$$

$$B_i(k) = \{b_i \in B_i(k-1) |\ b_i \in \Delta(\bigtimes_{j \neq i} (C_j \times B_j(k-1)))\},\ k > 1$$

A belief hierarchy b_i *expresses* **up to** k**-fold belief in the opponent's rationality** *if* $b_i \in B_i(k)$. *It expresses* **common belief in rationality** *if* $b_i \in B_i(\infty) = \bigcap_{k\geq 1} B_i(k)$.

From here, we straightforwardly introduce rational choice under belief in rationality at various levels:

Definition III.4. *(Rational Choice under k-Fold and Common Belief in Rationality)*
A choice c_i *for player* i *is*

a) **rational under up to** k**-fold belief in rationality** *for player* i *if there is a belief hierarchy* b_i *such that* c_i *is rational for* b_i *and* $b_i \in B_i(k)$.

b) **rational under common belief in rationality** *for player* i *if there is a belief hierarchy* b_i *such that* c_i *is rational for* b_i *and* $b_i \in B_i(\infty)$.

Like in traditional games, two questions about 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 – conditional on the existence of such belief hierarchies in a game Γ – there is an algorithm that allows us to find *all* choices for a player i that this player can make under common belief in rationality.
We investigate the first question in section IV and the second question in the remainder of the paper.

Before moving on, we also provide a definition of *psychological Nash equilibrium*, the most common solution concept used in the early literature on psychological games. As is already clear from Geanakoplos et al.'s (1989) original definition, psychological Nash equilibrium is characterized by common belief in rationality and the assumption that players' beliefs are generated by simple belief hierarchies (for an in-depth treatment for traditional games see Perea 2012). That is, there is common knowledge among players that a fixed probability distribution over all players' choices induces all players' beliefs. So, in particular, in a psychological Nash equilibrium players entertain *correct beliefs* about their opponents' beliefs about their opponents' choices, correct beliefs about the opponents' second-order beliefs, and so on.

To state this more formally, let $\sigma_i \in \Delta(C_i)$ be a probability distribution over i's choices and for a profile $\sigma = (\sigma_i)_{i\in I}$ of such distributions let $b_i[\sigma]$ be the belief hierarchy for player i where (1) i has belief σ_{-i} about the opponents' choices, (2) for every $j \neq i$, i assigns probability 1 to the event that j has belief σ_{-j} about the opponents' choices, and so on. We are now ready to define psychological Nash equilibrium:

Definition III.5. *(Psychological Nash Equilibrium)*
A vector of probability distributions $\sigma \in \times_{i\in I} \Delta(C_i)$ *is a* **psychological Nash equilibrium** *if, for every player* i *and every choice* $c_i \in \text{supp}(\sigma_i)$, *we have that* $u_i(c_i, b_i[\sigma]) \geq u_i(c_i', b_i[\sigma])$ *for all* $c_i' \in C_i$.

IV Possibility of Common Belief in Rationality

In this section we explore a condition, called *preservation of rationality at infinity*, which guarantees the existence of belief hierarchies expressing *common belief in rationality.* We start by defining this condition formally and show by means of a constructive proof that it always allows for belief hierarchies that express common belief in rationality. Subsequently, we show that this possibility result covers two interesting special cases: Games with utility functions that are *continuous at infinity*, and *belief-finite* games in which the utilities only depend on finitely many belief levels. Finally, we show by means of a counterexample that common belief in rationality may not be possible in games that do not preserve rationality at infinity.

A. 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 $(b_i^1, b_i^2, ...)$, where b_i^{n-1} and b_i^n always agree on the first $n-1$ orders of belief, then c_i must also be rational for the limit belief hierarchy it converges to.

Definition IV.1. *(Preservation of Rationality at Infinity)*
A psychological game $\Gamma = (C_i, B_i, u_i)_{i \in I}$ *is said to* **preserve rationality at infinity** *if for every player i, every choice $c_i \in C_i$ and every belief hierarchy $b_i \in B_i$, choice c_i is rational for b_i, if and only if, 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 $\hat{b}_i$.*

Equivalently, preservation of rationality at infinity states that 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 $\hat{b}_i^n = b_i^n$. An important difference relative to previously known existence results (cf. Geanakoplos et al. 1989, Battigalli and Dufwenberg 2009, and Bjorndahl et al. 2016) is that our proof is *constructive*, that is, we show how to *construct* a belief hierarchy expressing common belief in rationality under the assumption of preservation of rationality at infinity.

Theorem IV.2. *(Possibility of Common Belief in Rationality)*
Consider a psychological game $\Gamma = (C_i, B_i, u_i)_{i \in I}$ *that preserves rationality at infinity. Then, there is for every player i a belief hierarchy $b_i \in B_i$ that expresses common belief in rationality.*

Proof. For the proof we need a new piece of notation. Consider, for every $n \geq 1$, a choice combination $c^n = (c_i^n)_{i \in I}$ in $\times_{i \in I} C_i$. Then, we denote by $b_i[c^1, c^2, ...]$ 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 n-th order belief of $b_i[c^1, c^2, ...]$ by $(c^1, ..., c^n)$, and thus write $b_i^n[c^1, c^2, ...] = (c^1, ..., c^n)$.

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

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

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

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

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 $d(0) := (d_i(0))_{i \in I}$.

Then, for all players i, define a new belief hierarchy

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

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

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

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 combination $e^1 = (e_j^1)_{j \in I}$ in $\times_{j \in I} C_j$ 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 combination $e^2 = (e_j^2)_{j \in I}$ in $\times_{j \in I} C_j$ 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 combination $e^3 = (e_j^3)_{j \in I}$ in $\times_{j \in I} C_j$ such that there are infinitely many belief hierarchies $b_i \in \hat{B}_i[e^1, e^2]$ with $b_i^3 = (e^1, e^2, e^3)$. 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-combinations $e^1, e^2, ...,$ and we set

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

We now show that $\hat{b}_i$ expresses common belief in rationality. Suppose not. Then, there must be some $n \geq 1$ such that $\hat{b}_i$ does not express n-fold belief in rationality. Hence, there is some player j such that e_j^n is not rational for $b_j[e^{n+1}, e^{n+2}, ...]$. Since the game preserves rationality at infinity, there is some $k \geq 1$ such that e_j^n is not rational for any belief hierarchy b_j with

$$b_j^k = b_j^k[e^{n+1}, e^{n+2}, ...] = (e^{n+1}, ..., e^{n+k}).$$

As such, every belief hierarchy $b_i \in B_i$ with

$$b_i^{n+k} = \hat{b}_i^{n+k} = (e^1, ..., e^{n+k})$$

does not express n-fold belief in rationality.

Now, consider the infinite set $\hat{B}_i[e^1, ..., e^{n+k}]$. Since this set contains infinitely many belief hierarchies $b_i(m) \in \hat{B}_i$, and every $b_i(m)$ expresses up to m-fold belief in rationality, as we have seen, there must be some $b_i(m) \in \hat{B}_i[e^1, ..., e^{n+k}]$ with $m \geq n$ that expresses n-fold belief in rationality. As $b_i(m) \in \hat{B}_i[e^1, ..., e^{n+k}]$, we have that $b_i^{n+k}(m) = (e^1, ..., e^{n+k})$. Hence, we have found some belief hierarchy $b_i(m)$ with $b_i^{n+k}(m) = (e^1, ..., e^{n+k})$ that expresses n-fold belief in rationality. This, however, contradicts our finding above that there is no $b_i \in B_i$ with $b_i^{n+k} = \hat{b}_i^{n+k} = (e^1, ..., e^{n+k})$ that expresses n-fold belief in rationality.

Hence, we conclude that the belief hierarchy $\hat{b}_i$ must express 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. □

It is interesting to note that the construction performed in the proof of theorem IV.2 implies that in *all* psychological games (preserving rationality at infinity or not) we can find a belief hierarchy b_i for every player i such that b_i expresses up to k-fold belief in rationality for an arbitrary fixed $k \geq 1$. So up to k-fold belief in rationality can only ever fail at the limit where we try to extend a belief hierarchy expressing finitely many layers of belief in rationality to one that does so for all $k \in \mathbb{N}$.

As implied by theorem IV.2, it is not guaranteed that common belief in rationality is possible in games that do not preserve rationality at infinity. We will now present a concrete example of a game in which common belief in rationality is not possible:[2]

Example IV.3. (Common Belief in Rationality May not Be Possible)
Modified Bravery Game: (inspired by Geanakoplos et al. 1989)
Player 1 chooses to behave *timidly* or *boldly* while being observed by player 2. Player 1 is a timid guy so in almost all situations he prefers to behave timidly. Things are different, however, when he

[2]An example with the same structure, the *deeply surprising proposal*, has independently been developed by Bjorndahl et al. (2016).

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. In that case player 1 is angry and wants to prove player 2 wrong by choosing to act *boldly*.

Using the notation from the proof of theorem IV.2, let $b_1^{timid} = b_1[(timid, *), (timid, *), \ldots]$. In words, b_1^{timid} is 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".

Let the utility function for player 1 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 *timid* is always the unique rational choice for player 1, except when his belief hierarchy is b_1^{timid}. The game is summarized in table 1.

Table 1: Modified Bravery Game

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

Note that this game does not preserve rationality at infinity. Indeed, choice *timid* is not rational for the belief hierarchy b_1^{timid}, yet for every n we can find a belief hierarchy $\hat{b}_1$ with $\hat{b}_1^n = (b_1^{timid})^n$ such that *timid* is rational for $\hat{b}_1$.

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 player 2 believes that player 1 chooses *timid* and has belief hierarchy b_1^{timid}. However, *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. It follows that b_1^{timid} does not express up to 2-fold belief in rationality and, a fortiori, 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 the belief hierarchy 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 only choice *timid* is 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 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.*

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

B. Special Cases

We will now discuss some interesting special cases of psychological games that preserve rationality at infinity. First, we discuss games where the utility functions of the players are *continuous at infinity.* In such games, utility functions may depend on all orders of belief, but the impact that a specific order n has on the overall utility vanishes as n becomes large.

Definition IV.4. *(Continuity at Infinity)*
A psychological game $\Gamma = (C_i, B_i, u_i)_{i\in I}$ *is* **continuous at infinity** *if for every player* i*, every choice* c_i*, every belief hierarchy* b_i *and every* $\varepsilon > 0$ *there is some* $n \geq 1$ *such that for every belief hierarchy* $\hat{b}_i$ *with* $\hat{b}_i^n = b_i^n$ *we have that* $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$.

It may be verified that every game which is continuous at infinity will automatically preserve rationality at infinity. This is the content of the following lemma.

Lemma IV.5. *(Continuity at Infinity Refines Preservation of Rationality at Infinity)*
Every game that is continuous at infinity preserves rationality at infinity.

Proof. Consider a game $\Gamma = (C_i, B_i, u_i)_{i\in I}$ that is continuous at infinity, and take an arbitrary choice c_i and belief hierarchy b_i. Suppose that for every $n \geq 1$ there is some $b_i(n)$ with $b_i^n(n) = b_i^n$ such that c_i is rational for $b_i(n)$. We show that c_i is rational for b_i.

Suppose, contrary to what we want to show, that c_i is not rational for b_i. Then, there is some choice c_i' such that $u_i(c_i, b_i) < u_i(c_i', b_i)$. Define $\varepsilon := \frac{1}{2}(u_i(c_i', b_i) - u_i(c_i, b_i))$. Since the game is continuous at infinity, there is some $n \geq 1$ such that $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$ and $|u_i(c_i', b_i) -$

$u_i(c_i', \hat{b}_i)| < \varepsilon$ for every $\hat{b}_i$ with $\hat{b}_i^n = b_i^n$. In particular, it follows that

$$|u_i(c_i, b_i) - u_i(c_i, b_i(n))| < \varepsilon \quad \text{and} \quad |u_i(c_i', b_i) - u_i(c_i', b_i(n))| < \varepsilon.$$

Consequently,

$$\begin{aligned} u_i(c_i', b_i(n)) - u_i(c_i, b_i(n)) &= u_i(c_i', b_i) + (u_i(c_i', b_i(n)) - u_i(c_i', b_i)) \\ &\quad - u_i(c_i, b_i) - (u_i(c_i, b_i(n)) - u_i(c_i, b_i)) \\ &> u_i(c_i', b_i) - u_i(c_i, b_i) - 2\varepsilon = 0, \end{aligned}$$

which implies that c_i is not rational for $b_i(n)$. This, however, is a contradiction, and hence we conclude that c_i is rational for b_i. Therefore, the game preserves rationality at infinity. □

In view of Theorem IV.2 we may thus conclude that, in every game that is continuous at infinity, we always have belief hierarchies that express common belief in rationality.

We now turn to *belief-finite* games, where the utilities of the players only depend on finitely many orders of belief.

Definition IV.6. *(Belief-Finite Games)*
A psychological game $\Gamma = (C_i, B_i, u_i)_{i \in I}$ is **belief-finite** *if there is some $n \geq 1$ such that for every player i, every choice $c_i \in C_i$, and every two belief hierarchies b_i and $\hat{b}_i$ in B_i with $b_i^n = \hat{b}_i^n$ we have that $u_i(c_i, b_i) = u_i(c_i, \hat{b}_i)$.*

It is immediately clear that every belief-finite game is continuous at infinity, and hence, in view of Lemma IV.5, preserves rationality at infinity. This leads to the following observation.

Observation IV.7. *Every belief-finite game is continuous at infinity, and therefore preserves rationality at infinity.*

In view of Theorem IV.2 we may thus conclude that every belief-finite game allows for belief hierarchies that express common belief in rationality.

Together with our insights about games that are continuous at infinity, we arrive at the following conclusions.

Corollary IV.8. *(Special cases when common belief in rationality is possible)*
In every psychological game $\Gamma = (C_i, B_i, u_i)_{i \in I}$ that is belief-finite or continuous at infinity, we can always find for every player i a belief hierarchy b_i that expresses common belief in rationality.

C. Belief Continuity

Geanakoplos et al. (1989) show that for every psychological game with *continuous* utility functions given the product topology on B_i, we can always find a *psychological Nash equilibrium.* In what follows, we will refer to this continuity condition as *belief continuity.* Since a psychological Nash equilibrium is a special instance of a belief hierarchy expressing common belief in rationality, it follows from their result that common belief in rationality is always possible in a belief-continuous psychological game. In Battigalli and Dufwenberg (2009), we can also find a direct proof that belief continuity ensures that common belief in rationality is possible – not only for *static psychological games* as considered here but also for *dynamic psychological games.*[3]

In this section, we study how this "classical" existence condition can be characterized within our framework and how it relates to the sufficient conditions that we presented in the previous section. As our characterization shows, belief continuity implies *continuity at infinity.* We then move on to restate Geanakoplos et al.'s (1989) result regarding existence of psychological Nash equilibrium, accompanied by a novel example of a simple psychological game that allows for common belief in rationality but not for a psychological Nash equilibrium.

We start by defining belief continuity as introduced in Geanakoplos et al. (1989). To formally define this property, let $\hat{d}(b_i^k, \hat{b}_i^k)$ denote the Lévy-Prokhorov distance between two kth-order beliefs $b_i^k, \hat{b}_i^k \in B_i^k$ where $b_i^k, \hat{b}_i^k$ are viewed as probability measures on $C_{-i} \times B_{-i}^{k-1}$. Also, for belief hierarchies $b_i, \hat{b}_i \in B_i$, let $d(b_i, \hat{b}_i) = \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k \hat{d}(b_i^k, \hat{b}_i^k)$. It is well known that the distance d then metricizes the product space B_i. Given these preliminaries, we define:

Definition IV.9. *(Belief Continuity)*
A psychological game $\Gamma = (C_i, B_i, u_i)_{i \in I}$ *is* **belief-continuous** *if for every player* i*, every choice* c_i*, every belief hierarchy* b_i*, and every* $\varepsilon > 0$*, there is* $\delta > 0$ *such that for any belief hierarchy* $\hat{b}_i$ *with* $d(b_i, \hat{b}_i) < \delta$ *we have that* $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$*.*

Coming from *continuity at infinity* which allows for arbitrarily changing the belief hierarchy b_i at "high orders of belief", what we additionally allow for under *belief continuity* can be intuitively described as slightly changing the probabilistic n-th order belief b_i^n for a fixed n. The following characterization makes this precise:

Theorem IV.10. *(Robustness to Trembles of Finite Order Characterizes Belief Continuity)*
A psychological game $\Gamma = (C_i, B_i, u_i)_{i \in I}$ *is belief-continuous if and only if, for every player* i*, every choice* c_i*, every belief hierarchy* b_i*, and every* $\varepsilon > 0$*, there is* $k \in \mathbb{N}$ *and* $\delta > 0$ *such that for any belief* $\hat{b}_i$ *with* $\hat{d}(b_i^m, \hat{b}_i^m) < \delta$ *for all* $m \leq k$ *we have that* $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$*.*

[3] For dynamic games, Battigalli and Dufwenberg (2009) study common strong belief in rationality. So their existence result goes even a little farther in that they establish that also the existence of this refinement of common belief in rationality is always ensured under an appropriate generalization of belief continuity for dynamic games.

Proof.

$\Rightarrow$: To begin, assume that Γ is belief-continuous. Then, for all $c_i \in C_i$, $b_i \in B_i$ and $\varepsilon > 0$, there is $\delta > 0$ such that $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$ whenever $d(b_i, \hat{b}_i) < \delta$.

Now choose k such that $\sum_{m=k+1}^{\infty} \left(\frac{1}{2}\right)^m < \frac{\delta}{2}$. Further take $\hat{\delta} = \frac{\delta}{2}$ and let $\hat{b}_i \in B_i$ be such that $\hat{d}(b_i^m, \hat{b}_i^m) < \hat{\delta}$ for all $m \leq k$. Then

$$\begin{aligned} d(b_i, \hat{b}_i) &= \sum_{m=1}^{k} \left(\frac{1}{2}\right)^m \hat{d}(b_i^m, \hat{b}_i^m) + \sum_{m=k+1}^{\infty} \left(\frac{1}{2}\right)^m \hat{d}(b_i^m, \hat{b}_i^m) \\ &< \sum_{m=1}^{k} \left(\frac{1}{2}\right)^m \hat{\delta} + \sum_{m=k+1}^{\infty} \left(\frac{1}{2}\right)^m \\ &< \sum_{m=1}^{k} \left(\frac{1}{2}\right)^m \frac{\delta}{2} + \frac{\delta}{2} < \delta \end{aligned}$$

where for the first inequality we used $\hat{d}(b_i^m, \hat{b}_i^m) \leq 1$ for all $b_i^m, \hat{b}_i^m \in B_i^m$ and all $m \in \mathbb{N}$.

By definition of δ, it now follows that $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$, establishing the first direction.

$\Leftarrow$: Now assume Γ is such that, for every player i, every choice c_i, every belief hierarchy b_i, and every $\varepsilon > 0$, there is $k \in \mathbb{N}$ and $\delta > 0$ such that $\hat{d}(b_i^m, \hat{b}_i^m) < \delta$ for all $m \leq k$ implies $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$.

Now choose $\hat{\delta} = \frac{\delta}{2^k}$ and take $b_i, \hat{b}_i \in B_i$ such that $d(b_i, \hat{b}_i) < \hat{\delta}$.

Then

$$d(b_i, \hat{b}_i) = \sum_{m=1}^{k} \left(\frac{1}{2}\right)^m \hat{d}(b_i^m, \hat{b}_i^m) + \sum_{m=k+1}^{\infty} \left(\frac{1}{2}\right)^m \hat{d}(b_i^m, \hat{b}_i^m) < \frac{\delta}{2^k}.$$

So, in particular,

$$\sum_{m=1}^{k} \left(\frac{1}{2}\right)^m \hat{d}(b_i^m, \hat{b}_i^m) < \frac{\delta}{2^k}.$$

and hence $\hat{d}(b_i^m, \hat{b}_i^m) < \delta$ for all $m \leq k$.

By definition of δ, it now follows that $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$, establishing the second direction.

□

Clearly, if we demand the property from theorem IV.10 only for belief hierarchies that coincide on the first k levels, so that $\hat{d}(b_i^m, \hat{b}_i^m) = 0$ for all $m \leq k$, then we receive continuity at infinity. We therefore reach the following conclusion:

Corollary IV.11. *(Belief Continuity Refines Continuity at Infinity)*
If a game is belief-continuous, then it is continuous at infinity.

As is now clear, our existence condition preservation of rationality at infinity nests all previously known results regarding existence.

Geanakoplos et al. (1989) show that requiring belief continuity is even enough to prove the existence of a *psychological Nash equilibrium*:

Theorem IV.12. *(Existence of Psychological Nash Equilibrium)*
Let $\Gamma = (C_i, B_i, u_i)_{i \in I}$ be a belief-continuous psychological game. Then Γ has a psychological Nash equilibrium.

Proof. Shown in Geanakoplos et al. (1989), theorem 1. □

Building on this result and our lemma IV.5, the class of games that satisfy continuity at infinity but not belief continuity necessarily allows for common belief in rationality but does not necessarily allow for psychological Nash equilibrium. One prominent subclass of such games would be *belief-finite games* that are not *belief-continuous*. Below we provide an example of such a game that indeed has no psychological Nash equilibrium:

Example IV.13. (Non-Existence of Psychological Nash Equilibrium)
Modified Bravery Game II:
We consider a variation of the game from example IV.3. Different from before, player 1 already gets angry if he believes that player 2 *is sure* that player 1 will choose to behave timidly. In that case player 1 wants to prove player 2 wrong by choosing to act *boldly*.

Let $B_1(*, timid)$ bet the set of belief hierarchies for player 1 such that he believes that player 2 believes that he chooses *timid*. Here, "believes" means "assigns probability 1 to". The utility function for player 1 is now 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 $b_1 \in B_1(*, timid)$. That is, player 1 prefers to choose *bold* if and only if he is sure that player 2 believes him to choose *timid* with probability 1 and he prefers to choose *timid* otherwise. The game is summarized in table 2.

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

This game is trivially continuous at infinity as is any belief-finite psychological game. However, it is not belief-continuous since slightly perturbing second-order beliefs for any $b_1 \in B_1(*, timid)$ leads to discontinuous changes in $u_1(bold, b_1)$.
We now show that this game allows for common belief in rationality, but not for psychological Nash equilibrium.

To see that common belief in rationality is possible, we slightly vary the construction from the proof of theorem IV.2. Take a sequence of choice profiles $c^k \in C_1 \times C_2$ where $c^1 = c^2 = (timid, *)$, With the operator d from the proof of theorem IV.2 we now construct, for both players i,

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

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

$$b_i(k) = b_i[d(k), d(k-1), d(k-2), \dots]$$

Now note that $(d(k))_{k\in\mathbb{N}} = ((bold, *), (bold, *), (timid, *), (timid, *), (bold, *), (bold, *), \dots)$ such that the sequence of choice profiles enters a cycle.

This follows from the fact that *timid* is rational for player 1 whenever $b_1^2 = ((c_1, *), (bold, *)),\ c_1 \in \{bold, timid\}$ and that *bold* is rational for him whenever $b_1^2 = ((c_1, *), (timid, *)),\ c_1 \in \{bold, timid\}$. Since the belief hierarchies $\hat{b}_i = b_i[(bold, *), (bold, *), (timid, *), (timid, *), (bold, *), (bold, *), \dots]$ are hence generated by infinitely repeating cycles of choice profiles where each profile is rational given the second-order belief induced by the preceding two, we conclude that $\hat{b}_i$ expresses common belief in rationality for each player i.

To see that a psychological Nash equilibrium does not exist, note that there is no *simple* belief hierarchy that expresses common belief in rationality in this game. Since player 1 *strictly* prefers to choose *timid* whenever $b_1 \notin B_1(*, timid)$ and *strictly* prefers to choose *bold* otherwise, the only candidates for his equilibrium belief hierarchy would be $b_1[(timid, *), (timid, *), \dots]$ and $b_1[(bold, *), (bold, *), \dots)]$. However, since *timid* is not rational for player 1 if he entertains $b_1^2 = ((timid, *), (timid, *))$ and since *bold* is not rational for player 1 if he entertains $b_1^2 = ((bold, *), (bold, *))$, neither of these expresses up to 2-fold belief in rationality. It follows that there is no psychological Nash equilibrium in this game.

D. Elimination at the Limit

On top of not allowing for common belief in rationality, some psychological games can exhibit a second interesting peculiarity: Rational choice under common belief in rationality can *strictly refine* rational choice under up to k-fold belief in rationality for any finite k in these games. When that happens, we will be able to *eliminate choices at the limit* of common belief in rationality that can demonstrably be rationalized for any finite order of up to k-fold belief in rationality. This observation could already be made in example IV.3. Whereas the choice *timid* was rational for player 1 under up to k-fold belief in rationality for any k, no choice can be rationally made under common belief in rationality. In this section, we will provide a more positive example of a game in which common belief in rationality is possible, but where also elimination at the limit does occur.

Next, we provide a sufficient condition to ensure that elimination at the limit cannot occur in a psychological game. As it turns out, whenever a game is belief-continuous, any choice that can be made under k-fold belief in rationality for all finite $k \geq 1$ can also be made under common belief in rationality – revealing an additional property of belief continuity that was not known before.

To start, consider our example of a game that allows for common belief in rationality, but that also exhibits elimination at the limit:

Example IV.14. (Elimination of Choices at the Limit of Common Belief in Rationality)
Newcomb's Modified Paradox: (inspired by Nozick 1969)
A man can choose to take a *box* or that same box and a diamond (*both*). God can put *four* or *no* diamonds into the box. Different from the standard version of Newcomb's Paradox, god only wants to put no diamonds into the box if he considers it a *commonly known fact* that the man will go for *both*, not only believing that the man will choose *both*, but also believing that the man believes that he believes that the man will choose *both*, and so on. As in the standard case, the man simply chooses to maximize the number of diamonds he expects to receive.

Formally, we have a two-player game with player set $I = \{m, g\}$, choice set $C_m = \{box, both\}$ for man and choice set $C_g = \{four, no\}$ for god. Let $B_g(both)$ be the set of belief hierarchies for god where he believes that man chooses *both*, believes that man believes that god believes that man chooses *both*, believes that man believes that god believes that man believes that god believes that man chooses *both*, and so on. Again, "believes" means "assigns probability 1 to" here.

Let the utility function for man be $u_m(box, b_m) = 4b_m^1(four)$ and $u_m(both, b_m) = 1+4b_m^1(four)$. Hence, choice *both* strictly dominates choice *box* for man.

The utility function of god is given by $u_g(no, b_g) = 1$ and $u_g(four, b_g) = 0$ for $b_g \in B_g(both)$ and $u_g(no, b_g) = 0$ and $u_g(four, b_g) = 1$ for $b_g \notin B_g(both)$. This game is summarized in table 3 below.

Table 3: Newcomb's Modified Paradox

	c_g	
man	$four$	no
box	4	0
$both$	5	1

	b_g	
god	$b_g \in B_g(both)$	$b_g \notin B_g(both)$
$four$	0	1
no	1	0

Note that also this game does not preserve rationality at infinity: Choice $four$ is not rational for any belief hierarchy in $B_g(both)$. Yet for every n and every $b_g \in B_g(both)$ we can find a belief hierarchy $\hat{b}_g$ with $\hat{b}_g^n = b_g^n$ such that $\hat{b}_g \notin B_g(both)$, rendering $four$ rational under $\hat{b}_g$.

We first establish that god can choose $four$ under up to k-fold belief in rationality for any k. To prove this, we perform a similar construction as in the proof of theorem IV.2:

For $k \geq 1$ take a sequence of choice profiles $c^k \in C_m \times C_g$ and set $c^1 = (box, four)$. We use the operator d from the proof of theorem IV.2 to construct, for both players i,

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

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

$$b_i(k) = b_i[d(k), d(k-1), d(k-2), \dots, c^1, c^2, \dots]$$

Note that we have $d(k) = (both, four)$, $k \geq 1$ where $d_g(k) = four$ since $b_g(k) \notin B_g(both)$ by construction. Also, again by construction, $b_g(k)$ expresses up to k-fold belief in rationality.

Next, we show that god cannot choose $four$ under common belief in rationality:

Suppose that this were not the case. Then there must be a belief hierarchy $b_g \notin B_g(both)$ such that b_g expresses common belief in rationality. Since b_g must express up to 1-fold belief in rationality, it must be such that god believes that man chooses *both*. Also, since b_g must express up to 3-fold belief in rationality, it must be such that god believes that man believes that god believes that man chooses *both*. Continuing in this fashion for all odd levels k of up to k-fold belief in rationality, we conclude that $b_g \in B_g(both)$, a contradiction. We conclude that $four$ cannot be chosen under common belief in rationality, while it can be chosen under any finite level k of up to k-fold belief in rationality.

Example IV.14 shows that, in order to characterize all choices that can be made under common belief in rationality in a game that does not preserve rationality at infinity, we might have to exclude additional choices at the limit of common belief in rationality. Note that using the same technique as we used here, the reader may verify that, in example IV.3, player 1's choice *timid* can rationally be made under up to k-fold belief in rationality for every finite k whereas no choice can rationally be made under common belief in rationality. These examples are analogous to examples for traditional games presented in Lipman (1994), Dufwenberg and Stegemann (2002), and Bach and Cabessa (2012). Different from our examples IV.14 and IV.3, though, these traditional games exhibit *infinite* choice sets. We next investigate whether we can find a sufficient condition under which such an *elimination at the limit* cannot occur in a psychological game. As we now show, such a sufficient condition is provided by *belief continuity*:

Theorem IV.15. *(No Elimination at the Limit)*
Let $\Gamma = (C_i, B_i, u_i)_{i \in I}$ be a belief-continuous psychological game. Then whenever a choice $c_i \in C_i$ is rational for player i under up to k-fold belief in rationality for any $k \in \mathbb{N}$, it is also rational under common belief in rationality.

Proof. Assume that c_i is rational under up to k-fold belief in rationality for any $k \geq 0$ (where $k = 0$ is interpreted as rational choice). Let $B_i(k, c_i)$ be the set of belief hierarchies that rationalize c_i under up to k-fold belief in rationality. To prove our result, we show that $B_i(k, c_i)$ is a compact set for every $k \geq 0$. Since the sequence $B_i(0, c_i), B_i(1, c_i), \dots$ is then a decreasing sequence of nested compact sets, Cantor's intersection theorem implies that $B_i(\infty, c_i) = \bigcap_{k\geq 0} B_i(k, c_i)$ is non-empty such that c_i is indeed rational under 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 $b_j \notin B_j(0, c_j)$. 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 rational for 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 Polish, $B_j(0, c_j)$ is then also 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

$$\begin{aligned} B_i(k+1, c_i) &= \{b_i \in B_i(k, c_i) | b_i \in \Delta(\bigtimes_{j\neq i}\{(c_j, b_j) | c_j \in C_j,\ b_j \in B_j(k, c_j)\})\} \\ &= B_i(k, c_i) \cap \Delta(\bigtimes_{j\neq i}\{(c_j, b_j) | c_j \in C_j,\ b_j \in B_j(k, c_j)\}). \end{aligned}$$

By the induction assumption, every $B_j(k, c_j)$ is compact and metrizable such that $\times_{j\neq i}\{(c_j, b_j) | c_j \in C_j,\ b_j \in B_j(k, c_j)\}$ is compact and metrizable too. Since the set of probability measures over a compact and metrizable set is itself compact and metrizable, the same is true for $\Delta(\times_{j\neq i}\{(c_j, b_j) | c_j \in C_j,\ b_j \in B_j(k, c_j)\})$. It follows that $B_i(k+1, c_i)$ is compact and metrizable, completing the induction.

Cantor's intersection theorem now ensures that $B_i(\infty, c_i)$ is non-empty such that c_i is rational under common belief in rationality. □

V Common Belief in Rationality Characterized

In this section, we define an algorithm called *iterated elimination of choices and belief hierarchies* that characterizes common belief in rationality in general psychological games. The algorithm generalizes traditional *iterated elimination of strictly dominated choices* in an intuitive way. It proceeds by iterative elimination of combinations of choices and belief-hierarchies (c_i, b_i). At this point, it might not be all that obvious that we even have to generalize *iterated elimination of strictly dominated choices*, which characterizes common belief in rationality in traditional games, to tackle common belief in rationality in psychological games. Therefore, before introducing our algorithm, we present an example to convince ourselves that elimination of choices will not be enough to study common belief in rationality in most interesting psychological games. Subsequently, we formally define the algorithm, after which we illustrate it by means of an example.

A. Elimination of Choices is Not Enough

We will now discuss an example which shows that in a psychological game, elimination of choices alone may not be enough to arrive at the choices that can rationally be made under common belief in rationality.

Example V.1. (Elimination of Choices Does not Work in a Psychological Game)
Playing Hard to Get:
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 and *ditch* Alice. At the other end of town, Alice is asking herself the same question.

To have a good evening no matter what, you suggested your favorite bar, so already without the date you prefer not to stay home. Obviously, though, you still like it more if Alice comes than otherwise.

At the same time, Alice seemed very confident that you would want to date her if only she agrees and you are still a bit annoyed by that fact. That is why you also consider ditching her. You would enjoy that if you think she is very sure that you go to the bar. Since you then ditched her, it is clear that there will not be another date. So you do not care what she does in this case.

Alice's preferences are less capricious. She prefers to go if she thinks you will likely come and otherwise she prefers to ditch you.

Formally, this is a two-player psychological game Γ in which $I = \{y, a\}$ and $C_y = C_a = \{date, ditch\}$. No different from a traditional game, Alice's utility function only depends on first-order beliefs. Specifically:

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

Different from a traditional game, the utility function of you depends on both first- and second-order beliefs. Let it be defined as follows

$$u_y(date, b_y) = 1 + b_y^1(date),\ u_y(ditch, b_y) = \int_{C_a \times B_a} b_a^1(date)\, \mathrm{d}b_y =: \varepsilon_y^2(date)$$

Here $\varepsilon_y^2(date)$ represents the expected probability you think Alice assigns to your choice *date*.

As implied by a more general characterization theorem for *additive* games in our companion paper Jagau and Perea (2017), Alice's and your utility functions can be summarized by one finite matrix for Alice and two finite matrices for you as shown in table 4 below.

Alice's matrix and your first matrix collect the utility the given player derives from probability 1 first-order beliefs. The second matrix for you collects the utility that depends on your second-order beliefs. More precisely, only the *expected probability* which you believe Alice to assign to you choosing *date* matters for your utility. We call this your *second-order expectation* regarding

Table 4: Playing Hard to Get

b_y^1

You	*date*	*ditch*
date	2	1
ditch	0	0

$+$

ε_y^2

You	*date*	*ditch*
date	0	0
ditch	1	0

b_a^1

Alice	*date*	*ditch*
date	1	0
ditch	0	1

your choice *date*. As implied by the mentioned representation theorem, it is enough to collect the utility you derive from the extreme second-order expectations ($\varepsilon_y^2(date) = 1$ and $\varepsilon_y^2(date) = 0$) in the second matrix.

The resulting psychological game is about as well-behaved as a psychological game can be without being a traditional game.[4] Still, already for this game, we can show that iterated elimination of strictly dominated choices, which characterizes common belief in rationality in any traditional game, will not suffice to characterize common belief in rationality in Playing Hard to Get.
To see this, first note that every choice in this game 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 *ditch* is rational otherwise.
- For you, choosing *ditch* is rational whenever you believe, with probability 1, that Alice chooses *ditch* and believes, again with probability 1, that you choose *date*. For any other belief of you, your choice *date* is rational.

Since any choice of any player can be 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.

The reasoning goes as follows: Given the coordinative nature of Alice's decision problem, she should choose *date* when she deems it more likely that you choose *date* and she should choose *ditch* otherwise. At the same time, you can only choose *ditch* if you are sure that Alice chooses *ditch* and thinks that you choose *date*. However, in this case, you would not believe in Alice's rationality.

[4]The reader may check that in terms of classifications in section VII and in our companion paper Jagau and Perea (2017), playing hard to get is both a unilateral and an expectation-based psychological game which can be shown to imply that common belief in rationality for this game can be characterized by a *finite algorithm* that proceeds by iteratively imposing *linear restrictions* on choices and beliefs.

Hence, under common belief in rationality you can only choose *date*. As such, also Alice can only choose *date* under common belief in rationality.

Note that, different from what we can observe in traditional games, there are no irrational choices for any player in Playing Hard to Get, but there is a choice, namely your choice *ditch*, 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. This is precisely the reason why iterated elimination of choices does not characterize common belief in rationality in Playing Hard to Get.

B. Iterated Elimination of Choices and Belief Hierarchies

It might now seem clear that in a general psychological game, where utilities may depend non-linearly on arbitrary levels of beliefs, we cannot even do better than directly eliminating in the full belief space. In fact, our analysis of example IV.3 already shows that we sometimes need to rely on *all* information encoded in a belief hierarchy b_i to determine which choices are rational under common belief in rationality for a given player in a psychological game.[5] One might expect that things get simpler in a belief-finite game. In section VI we will show that this assumption indeed allows for a simpler elimination procedure.

Procedure V.2. *(Iterated Elimination of Choices and Belief Hierarchies)*

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

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

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

$$R_i(k) = \{(c_i, b_i) \in R_i(k-1) |\ b_i \in \Delta(R_{-i}(k-1))\}.$$

We finally define:

$$R_i(\infty) = \bigcap_{k \geq 1} R_i(k).$$

We collect the basic properties of iterated elimination of choices and belief hierarchies in the following observation:

[5] More formally, this is seen in example V.4 below where we apply iterated elimination of choices and belief hierarchies to the game from example IV.3.

Observation V.3. *(Basic Properties of Iterated Elimination of Choices and Belief Hierarchies)*

a) *For every k, the belief hierarchies b_i that exhibit up to k-fold belief in rationality are exactly the belief hierarchies surviving $k+1$ consecutive steps of elimination of choices and belief hierarchies. Also the choices that can be made under up to k-fold belief in rationality are exactly the choices in the projection* $\mathrm{proj}_{C_i}(R_i(k+1))$.

b) *The belief hierarchies b_i that exhibit common belief in rationality, if existent, are exactly the belief hierarchies that survive iterated elimination of choices and belief hierarchies. The choices that can be rationally made under common belief in rationality are exactly the choices in the projection* $\mathrm{proj}_{C_i}(R_i(\infty))$.

We refrain from proving the equivalence of $k+1$ steps of iterated elimination of choices and belief hierarchies and up to k-fold belief in rationality as introduced in definition III.3 since it is obvious from inspection: For each layer k of belief in rationality, we there require that types only deem possible opponents' belief hierarchies that express up to $k-1$-fold belief in rationality. That is exactly what we are doing if, moving from $R_i(k)$ to $R_i(k+1)$, we require that player i's belief hierarchy should be induced by a probability distribution over combinations of choices and belief hierarchies in $\Delta(R_{-i}(k))$.

C. Example

At every step, the procedure restricts *choice-belief combinations* according to increasing layers of belief in the opponents' rationality. For traditional games, this will yield the same reduction of the choice set as iterated elimination of strictly dominated choices would. However, the output of the procedure lives in $\times_{i\in I}(C_i \times B_i)$, so it actually corresponds to an epistemic model. Specifically, it corresponds to the largest epistemic model for the given game where all types express common belief in rationality. In traditional games we can, but do not have to, drag along all this additional information while iteratively characterizing common belief in rationality. In contrast, the potential dependence of utility on the full belief hierarchy in psychological games does not in general allow us to disregard any part of the information encoded in the space of belief hierarchies at any step of the iterative characterization.

By observation V.3, iterated elimination of choices and belief hierarchies characterizes common belief in rationality for any psychological game. In particular, we should then expect that the procedure yields an empty reduction when applied to the game from example IV.3. As we now show, this is indeed the case.[6]

[6] In a similar fashion, we can use iterated elimination of choices and belief hierarchies to show that $four$ gets eliminated at the limit of common belief in rationality in Newcomb's Modified Paradox as introduced in example IV.14.

Example V.4. (The Procedure when Common Belief in Rationality Is not Possible) Reconsider the Modified Bravery Game from example IV.3. We will now apply iterated elimination of choices and belief hierarchies to this game. Before we start, it is useful to define, for $n \geq 1$ and b_1^{timid}, the belief hierarchy we singled out earlier,

$$B_1^{(n)}(b_1^{timid}) = \{b_1 \in B_1 | b_1^n = b_1^{timid,n}\}$$

– the set of belief hierarchies for player 1 that induce the same nth-order belief as b_1^{timid}.

We also define $B_1^{(0)}(b_1^{timid}) = B_1$ and note that $\bigcap_{n\in\mathbb{N}} B_1^{(n)}(b_1^{timid}) = \{b_1^{timid}\}$. Given these preliminaries, the procedure yields

1. $R_1(1) = \{(bold, b_1^{timid})\} \cup \{(timid, b_1) | b_1 \neq b_1^{timid}\}$ and $R_2(1) = C_2 \times B_2$
2. $R_1(2) = R_1(1)$ and $R_2(2) = \{(*, b_2) | b_2 \in \Delta(\{(bold, b_1^{timid})\} \cup \{(timid, b_1)|b_1 \neq b_1^{timid}\})\}$
3. $R_1(3) \subseteq \{(timid, b_1) | b_1 \neq b_1^{timid}\}$ and $R_2(3) = R_2(2)$
4. $R_1(4) = R_1(3)$ and $R_2(4) \subseteq \{(*, b_2) | b_2 \in \Delta(\{(timid, b_1) | b_1 \neq b_1^{timid}\})\}$
5. $R_1(5) \subseteq \{(timid, b_1) | b_1 \in B_1^{(2)}(b_1^{timid}) \backslash \{b_1^{timid}\}\}$ and $R_2(5) = R_2(4)$

Continuing in this fashion we obtain, for any $k \geq 0$:

$$R_1(3+2k) \subseteq \{(timid, b_1) | b_1 \in B_1^{(2k)}(b_1^{timid}) \backslash \{b_1^{timid}\}\} \text{ and}$$

$$R_2(4+2k) \subseteq \{(\star, b_2) | b_2 \in \Delta(\{(timid, b_1) | b_1 \in B_1^{(2k)}(b_1^{timid}) \backslash \{b_1^{timid}\}\})\}.$$

In the limit we obtain

$$\bigcap_{k\in\mathbb{N}} R_1(k) \subseteq \emptyset \text{ and}$$

$$\bigcap_{k\in\mathbb{N}} R_2(k) \subseteq \emptyset.$$

In agreement with our impossibility result from example IV.3, iterated elimination of choices and belief hierarchies yields an empty reduction. Note that $R_i(k) \neq \emptyset$, $i \in \{1,2\}$ for any finite k.[7] So we need to use *all information* encoded in players' belief hierarchies to determine the (empty) set of combinations of choices and belief hierarchies expressing common belief in rationality in this game.

As the example shows, the general way to eliminating choices that are inconsistent with common belief in rationality in psychological games can be much more intricate than in traditional games, partly because we need to drag along much more information about players' beliefs than for standard elimination procedures.

[7] While it is not straightforward to write down the exact reduction generated by the procedure, we can easily construct belief hierarchies consistent with up to k-fold belief in rationality using the method from theorem IV.2 which suffices to show that each finite reduction is non-empty.

In the remainder of this paper, we will consider conditions under which elimination of choices and belief hierarchies can be replaced by a simpler procedure that keeps track of less information about belief hierarchies.

VI Belief-Finite Games

In this section we introduce a procedure, *Iterated Elimination of choices and nth-order beliefs*, that simplifies elimination of choices and belief hierarchies by treating belief hierarchies that coincide in the nth-order beliefs equally.[8] We will show that $k+1$ steps of this procedure characterize up to k-fold belief in rationality in every belief-finite psychological game where utilities depend on at most $n+1$ levels of beliefs. Further, we prove that the characterization goes through to common belief in rationality if the game, in addition, is belief-continuous. For the non-belief-continuous case, we present a slightly more involved procedure that provides an exact characterization. Subsequently, we illustrate *Iterated Elimination of choices and nth-order beliefs* by means of an example.

A. Iterated Elimination of Choices and nth-Order Beliefs

Henceforth, we will assume that utility functions only depend on $n+1$th-order beliefs so that we can write utilities as functions

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

Procedure VI.1. *(Iterated Elimination of Choices and nth-Order Beliefs)*

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

$$\begin{aligned} R_i^n(1) = \{ &(c_i, b_i^n) \in C_i \times B_i^n | \exists b_i^{n+1} \in B_i^{n+1} \textit{ with } \mathrm{marg}_{X_i^n} b_i^{n+1} = b_i^n \\ &\textit{such that } u_i(c_i, b_i^{n+1}) \geq u_i(c_i', b_i^{n+1}), \forall c_i' \in C_i \}. \end{aligned}$$

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

$$\begin{aligned} R_i^n(k) = \{ &(c_i, b_i^n) \in R_i^n(k-1) | \exists b_i^{n+1} \in \Delta(R_{-i}^n(k-1)) \textit{ with } \mathrm{marg}_{X_i^n} b_i^{n+1} = b_i^n \\ &\textit{such that } u_i(c_i, b_i^{n+1}) \geq u_i(c_i', b_i^{n+1}), \forall c_i' \in C_i \}. \end{aligned}$$

We finally define:

$$R_i^n(\infty) = \bigcap_{k \geq 1} R_i^n(k).$$

[8] For traditional games, the algorithm *iterated elimination of choices and 0th-order beliefs* defined in this section is exactly the same as iterated elimination of strictly dominated choices.

Elimination of choices and nth-order beliefs coincides with the full-blown elimination of choices and belief hierarchies except for keeping track only of nth-order beliefs. In a psychological game where utilities depend only on up to $n+1$th-order beliefs, this is enough to iteratively characterize all choices and all nth-order beliefs that are consistent with k-fold-belief in rationality. Also, any combination of choices and nth-order beliefs that is consistent with common belief in rationality does survive the procedure. Conversely, due to possible elimination at the limit, it is not necessarily true that any combination of choices and nth-order beliefs that survives the procedure is consistent with common belief in rationality. This *will* be true, however, provided that the game under study is belief-continuous. We establish all these results in the next theorem. To state the theorem compactly, we define:

Definition VI.2. *(Consistency with up to k-Fold and Common Belief in Rationality)*
A choice-belief combination $(c_i, b_i^n) \in C_i \times B_i^n$ *for player* i *is*

a) **consistent with up to k-fold belief in rationality** *for player* i *if there exists a belief hierarchy* b_i *that expresses up to* k*-fold belief in rationality, induces* b_i^n*, and rationalizes* c_i*.*

b) **consistent with common belief in rationality** *for player* i *if there exists a belief hierarchy* b_i *that expresses common belief in rationality, induces* b_i^n*, and rationalizes* c_i*.*

We are now ready to state theorem VI.3:

Theorem VI.3. *(The Algorithm Works)*
Take a psychological game Γ *in which utilities depend only on* $n+1$*th-order beliefs.*

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. *Any choice-belief combination* $(c_i, b_i^n) \in C_i \times B_i^n$ *that is consistent with common belief in rationality is in* $R_i^n(\infty)$*.*

3. *In a belief-continuous game, any choice-belief combination* (c_i, b_i^n) *in* $R_i^n(\infty)$ *is consistent with common belief in rationality.*

Proof.

Part 1:

$\Rightarrow$ 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: Suppose that (c_i, b_i^n) is consistent with 0-fold belief in rationality. Then c_i is rational for some belief hierarchy b_i that induces b_i^n. Since utility depends on at most $n+1$ belief levels, the $n+1$th-order belief b_i^{n+1} that is 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 $(c_i, b_i^n) \in R_i^n(1)$ since $b_i^n = \mathrm{marg}_{X_i^n} b_i^{n+1}$.

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 some $b_i \in B_i$ 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 combinations (c_{-i}, b_{-i}) of opponents' choices and belief hierarchies, 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. $b_i^n = \text{marg}_{X_i^n} b_i^{n+1}$.

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

$\Leftarrow$ 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, completing the induction start.

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 $(c_j, b_j^n) \in \Theta_j^n$, there is a belief hierarchy $b_j =: \theta_j(c_j, b_j^n)$ that expresses up to k-fold belief in rationality, induces b_j^n and rationalizes c_j. Given the mapping θ_j, for any measurable $E_j^n \subseteq \Theta_j^n$, let $\theta_j(E_j) = \{\theta_j(c_j, b_j^n) | (c_j, b_j^n) \in \Theta_j^n\}$. Now let b_i be the belief hierarchy given by $b_i^{n+1}(E_{-i}^n) = b_i(\times_{j \neq i} \theta_j(E_j^n))$ for every measurable $E_{-i}^n \subseteq \times_{j \neq i} \Theta_j^n$. Since b_i

assigns probability 1 to combinations of choices and belief hierarchies $(c_j, b_j) = (c_j, \theta_j(c_j, b_j^n))$ such that $\theta_j(c_j, b_j^n)$ expresses up to k-fold belief in rationality and rationalizes c_j, it follows that b_i expresses up to $k+1$-fold belief in rationality. Moreover, 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 of part 1.

Part 2:
Part 2 directly follows from part 1 and the fact that any choice-belief combination (c_i, b_i^n) that is consistent with common belief in rationality is automatically consistent with up to k-fold belief in rationality for any $k \geq 0$. So any (c_i, b_i^n) that is consistent with common belief in rationality will certainly survive the algorithm.

Part 3:
Part 3 emerges as a consequence of part 1 and our theorem IV.15: Take a belief-finite game in which utilities depend on at most $n+1$th-order beliefs and, furthermore, assume that the game is belief-continuous. Let $(c_i, b_i^n) \in R_i^n(\infty)$. Again, by part 1, there is a sequence $(b_i(k))_{k\in\mathbb{N}}$ of belief hierarchies, where each $b_i(k)$ induces b_i^n, expresses up to k-fold belief in rationality, and rationalizes c_i. Hence $b_i(k) \in B_i(k, c_i)$ for every k where $B_i(k, c_i)$ is defined as in the proof of theorem IV.15. Since B_i is Polish and thereby sequentially compact, $(b_i(k))_{k\in\mathbb{N}}$ has a converging subsequence $(b_i'(k))_{k\in\mathbb{N}}$, the limit of which we denote by $b_i'(\infty)$. Note that, clearly, $b_i'(\infty)$ induces b_i^n. Now, as we saw in the proof of theorem IV.15, $B_i(k, c_i)$ is compact for every $k \geq 1$. So fix some arbitrary k. Then $b_i'(m) \in B_i(k, c_i)$ for all $m \geq k$ and $b_i'(\infty) \in B_i(k, c_i)$ by compactness of $B_i(k, c_i)$. Since k was arbitrary, we can thus conclude that $b_i'(\infty) \in B_i(c_i, \infty)$. Since $b_i'(\infty)$ induces b_i^n, it follows that (c_i, b_i^n) is consistent with common belief in rationality. □

If a belief-finite game is *not* belief-continuous, iterated elimination of choices and nth-order beliefs will in general not provide an exact characterization of common belief in rationality. The reason is that we might have to reckon with elimination at the limit: If players' utilities depend on $n+1$th-order beliefs in a non-belief-continuous game, then we might have a choice-belief combination (c_i, b_i^n) that can be rationalized under up to k-fold belief in rationality using some belief hierarchy $b_i(k)$ for any given k, but, since the reductions $R_i^n(k)$ are not necessarily closed sets, it might be that none of these belief hierarchies does the trick *for all k at the same time.* Then (c_i, b_i^n) would end up in $R_i^n(\infty)$, but clearly it would not be consistent with common belief in rationality. So surviving iterated elimination of choices and nth-order beliefs is only a *necessary*, and not a *sufficient*, condition for a choice-belief combination to be consistent with common belief in rationality in such games.

We will now introduce a procedure, called *iterated elimination of choices and nth- and higher-order beliefs* that *does* exactly characterize common belief in rationality in belief-finite belief-discontinuous games while using strictly less information than we would use under iterated elimination of choices and belief hierarchies. That procedure, however, is substantively more complicated than iterated elimination of choices and nth-order beliefs. At any given step k of the procedure, instead of tracing beliefs up to the *penultimate utility-relevant* level, we will need to trace them up to the *ultimate* level *relevant for up to* $k-1$*-fold belief in rationality.* For notational convenience let utility functions only depend on nth-order beliefs in what follows, so that we can write

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

Procedure VI.4. *(Iterated Elimination of Choices and nth- and Higher-Order Beliefs)*

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

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

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

$$R_i^{n\uparrow}(k) = \{(c_i, b_i^{n+(k-1)}) \in C_i \times B_i^{n+(k-1)} | (c_i, b_i^{n+(k-2)}) \in R_i^{n\uparrow}(k-1),\ b_i^{n+(k-1)} \in \Delta(R_{-i}^{n\uparrow}(k-1))\}.$$

Let $\overline{R}_i(k) = \{(c_i, b_i) \in C_i \times B_i | (c_i, b_i^{n+(k-1)}) \in R_i^{n\uparrow}(k)\}$. We finally define:

$$R_i^{n\uparrow}(\infty) = \bigcap_{k \geq 1} \overline{R}_i(k).$$

We will now prove the following result:

Theorem VI.5. *(The Algorithm Works)*
Take a psychological game Γ in which utilities depend only on nth-order beliefs. The choice-belief combinations (c_i, b_i^n) that are consistent with common belief in rationality are exactly the choice-belief combinations in $\mathrm{proj}_{C_i \times B_i^n} R_i^{n\uparrow}(\infty)$.

Proof.
To prove the statement, we show that $R_i^{n\uparrow}(k) = R_i^{n+(k-1)}(k)$ and $\overline{R}_i(k) = R_i(k)$ for all $k \in \mathbb{N}$ and all players i. Here $R_i^{n+(k-1)}(k)$ is the reduction generated by *iterated elimination of choices and $n+(k-1)$th-order beliefs* (cf. procedure VI.1) and $R_i(k)$ is the reduction generated by *iterated elimination of choices and belief hierarchies* (cf. procedure V.2). The characterization then directly follows from the definition of procedure V.2. Note that $\overline{R}_i(k) = R_i(k)$ also implies $R_i^{n+k+m}(k) = \{(c_i, b_i^{n+k+m}) \in C_i \times B_i^{n+k+m} | (c_i, b_i^{n+(k-1)}) \in R_i^{n\uparrow}(k)\}$ for all $k \in \mathbb{N}$, all $m \in \mathbb{N}_0$, and all players i. In words, if we can complete $R_i^{n\uparrow}(k)$ in a way that yields $R_i(k)$, then we can use the same technique to receive any intermediate-size reduction $R_i^{n+k+m}(k)$. This fact will be used extensively below. We prove the statement that $R_i^{n\uparrow}(k) = R_i^{n+(k-1)}(k)$ and $\overline{R}_i(k) = R_i(k)$ by induction over $k \geq 1$:

Induction Start: For $k = 1$, the statement follows directly from the fact that utilities depend on at most nth-order beliefs.

Induction Step: Assume that, indeed, $R_i^{n\uparrow}(k) = R_i^{n+(k-1)}(k)$ and $R_i(k) = \overline{R}_i(k)$ for $k \geq 1$, and all players i. Then

$$\begin{aligned}
R_i^{n+k}(k+1) =& \{(c_i, b_i^{n+k}) \in R_i^{n+k}(k) | \exists b_i^{n+k+1} \in \Delta(R_{-i}^{n+k}(k)) \text{ with } \mathrm{marg}_{X_i^{n+k}} b_i^{n+k+1} = b_i^{n+k} \\
& \quad \text{such that } u_i(c_i, b_i^{n+k+1}) \geq u_i(c_i', b_i^{n+k+1}), \forall c_i' \in C_i\} \\
=& \{(c_i, b_i^{n+k}) \in C_i \times B_i^{n+k} | (c_i, b_i^{n+k-1}) \in R_i^{n\uparrow}(k) \\
& \quad \text{and } \exists b_i^{n+k+1} \in \Delta(R_{-i}^{n+k}(k)) \text{ with } \mathrm{marg}_{X_i^{n+k}} b_i^{n+k+1} = b_i^{n+k} \\
& \quad \text{such that } u_i(c_i, b_i^{n}) \geq u_i(c_i', b_i^{n}), \forall c_i' \in C_i\} \\
=& \{(c_i, b_i^{n+k}) \in C_i \times B_i^{n+k} | (c_i, b_i^{n+(k-1)}) \in R_i^{n\uparrow}(k) \\
& \quad \text{and } \exists b_i^{n+k+1} \in \Delta(\bigtimes_{j \neq i} \{(c_j, b_j^{n+k}) \in C_j \times B_j^{n+k} | (c_j, b_j^{n+(k-1)}) \in R_j^{n\uparrow}(k)\}) \\
& \quad \text{with } \mathrm{marg}_{X_i^{n+k}} b_i^{n+k+1} = b_i^{n+k} \text{ such that } u_i(c_i, b_i^{n}) \geq u_i(c_i', b_i^{n}), \forall c_i' \in C_i\} \\
=& \{(c_i, b_i^{n+k}) \in C_i \times B_i^{n+k} | (c_i, b_i^{n+(k-1)}) \in R_i^{n\uparrow}(k),\ b_i^{n+k} \in \Delta(R_{-i}^{n\uparrow}(k))\} \\
=& R_i^{n\uparrow}(k+1).
\end{aligned}$$

Here, for the second and third equality, we used that $R_i^{n+k}(k) = \{(c_i, b_i^{n+k}) \in C_i \times B_i^{n+k} | (c_i, b_i^{n+(k-1)}) \in R_i^{n+(k-1)}(k)\} = \{(c_i, b_i^{n+k}) \in C_i \times B_i^{n+k} | (c_i, b_i^{n+(k-1)}) \in R_i^{n\uparrow}(k)\}$ for all players i. This establishes the first statement.

Further, we have

$$\begin{aligned}
R_i(k+1) =& \{(c_i, b_i) \in R_i(k) | b_i \in \Delta(R_{-i}(k))\} \\
=& \{(c_i, b_i) \in \overline{R}_i(k) | b_i \in \Delta(\overline{R}_{-i}(k))\} \\
=& \{(c_i, b_i) \in C_i \times B_i | (c_i, b_i^{n+(k-1)}) \in R_i^{n\uparrow}(k) \text{ and } b_i \in \Delta(\overline{R}_{-i}(k))\} \\
=& \{(c_i, b_i) \in C_i \times B_i | (c_i, b_i^{n+(k-1)}) \in R_i^{n\uparrow}(k) \\
& \quad \text{and } b_i \in \Delta(\bigtimes_{j \neq i} \{(c_j, b_j) \in C_j \times B_j | (c_j, b_j^{n+(k-1)}) \in R_j^{n\uparrow}(k)\})\} \\
=& \{(c_i, b_i) \in C_i \times B_i | (c_i, b_i^{n+(k-1)}) \in R_i^{n\uparrow}(k) \text{ and } b_i^{n+k} \in \Delta(R_{-i}^{n\uparrow}(k))\} \\
=& \{(c_i, b_i) \in C_i \times B_i | (c_i, b_i^{n+k}) \in R_i^{n\uparrow}(k+1)\} \\
=& \overline{R}_i(k+1)
\end{aligned}$$

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

B. Example

We illustrate iterated elimination of choices and nth-order beliefs using the game introduced in example V.1:

Example VI.6. (The Procedure in Playing Hard to Get)
In this example, we reconsider Playing Hard to Get as first discussed in example V.1. Since all players' utilities in this game depend only on second-order beliefs and since the game is belief-continuous, we can apply iterated elimination of choices and 1st-order beliefs to determine the choice-belief combinations that are consistent with common belief in rationality. We proceed as follows:

1. $R^1_y(1) = \{(ditch, b^1_y)|b^1_y(ditch) = 1\} \cup \{(date, b^1_y)|b_y \in B^1_y\}$ and
   $R^1_a(1) = \{(date, b^1_a)|b^1_a(date) \geq \frac{1}{2}\} \cup \{(ditch, b^1_a)|b^1_a(date) \leq \frac{1}{2}\}$.
2. $R^1_y(2) = \{(date, b^1_y)|b^1_y \in B^1_y\}$ and $R^1_a(2) = R^1_a(1)$.
3. $R^1_y(3) = R^1_y(2)$ and $R^1_a(3) = \{(date, b^1_a)|b^1_a(date) = 1\} =: \{(date, date)\}$.
4. $R^1_y(4) = \{(date, b^1_y)|b^1_y(date) = 1\} =: \{(date, date)\}$ and $R^1_a(4) = R^1_a(3)$.

After four steps of elimination, only a unique combination of choices and first-order beliefs remains admissible for each player so that the procedure has converged. It follows that $(date, date)$ is the only choice-belief combination that is consistent with common belief in rationality for both you and Alice. Note that, different from what we can observe under regular iterated elimination of dominated choices, elimination of choices under the present procedure kicks in at the second step only and three steps of the procedure select *date* as the unique choice that is consistent with common belief in rationality for both you and Alice .

This mirrors the fact, mentioned earlier, that there are no irrational choices in Playing Hard to Get, but there is a choice, namely your choice *ditch*, that is not rational under 1-fold belief in rationality.

Even though keeping track of a finite number of payoff-relevant belief-levels considerably simplifies things, elimination of choices and nth-order beliefs can still take an infinite number of steps to converge for suitably specified utility functions as will be illustrated in the next subsection. So we will need to restrict admissible utility functions if we want to end up with a finite elimination procedure. One way in which we can do this will be explored in section VII.

C. The Procedure is Not Finite

We will now show by means of an example that already in the simplest non-degenerate case of a 2×2-psychological game where both players only care about the first- and second-order beliefs,[9] the procedure does not necessarily terminate within finitely many steps.

Example VI.7. (Procedure May Not Terminate within Finitely Many Steps)
The Nightly Encounter:
Going home after another evening in your favorite bar, Alice and you are shortcutting through a back-alley when, suddenly, a menacing figure appears from out of the shadows. Both Alice and you must think quickly, you can either *stay* or *run*.

Clearly you would never want to run and leave Alice behind or to be left behind by her. At the same time, you have a pretty bad feeling about the situation so you would prefer both of you just running for it to staying and facing the potential danger together. In addition, you care about what Alice expects you to do. In particular, if she believes that you will run anyway, then you hate the idea of playing the bold guy and staying. At the same time, if she expects you to be bold then you do not want to be the coward that ends up running away. Since deep inside you are still uncomfortable with the thought of staying in the first place, you like it better to run away when Alice expects you to than you like it to stay when Alice expects that.

Alice's preferences are similar to yours: She also would always rather have you both run or stay than having one of you being left behind by the other. Also she does not like the idea of playing bold when you expect her to make a run or of running away when you expect her to be bold. However, she is less terrified by the menacing figure than you are, so that she tends to think that running away would be unnecessarily cautious.

We model this situation as a 2×2-psychological game with player set $I = \{y, a\}$ and choice sets $C_y = C_a = \{stay,\ run\}$. Let your utility function be given by

$$u_y(stay, b_y) = 2(b_y^1(stay) + \varepsilon_y^2(stay)) \text{ and } u_y(run, b_y) = 3(b_y^1(run) + \varepsilon_y^2(run)).$$

Similarly, Alice's utility function is given by

$$u_a(stay, b_a) = 3(b_a^1(stay) + \varepsilon_a^2(stay)) \text{ and } u_a(run, b_a) = 2(b_a^1(run) + \varepsilon_a^2(run)).$$

As in example V.1, we define $\varepsilon_i^2(c_i) = \int_{C_j \times B_j} b_j^1(c_i)\, \mathrm{d}b_i$ for $i \in \{a, y\}$. Recall that this expression, which we again refer to as the *second-order expectation* of player i regarding c_i, captures the *expected probability* which i believes the opponent to assign his choice c_i. As implied by our representation theorem in Jagau and Perea (2017), we can represent Alice's and your preferences by two pairs

[9] In fact, the game we discuss here has especially nice properties in that it is also *additive*. As we will see, common belief in rationality can here be characterized by an LP-implementable algorithm. This applies more generally in additive games, cf. our results in Jagau and Perea (2017).

of finite matrices containing the utilities that you and Alice derive from your extreme first-order beliefs and your extreme second-order expectations. This is shown in table 5 below. The total

Table 5: The Nightly Encounter

	b_y^1				ε_y^2	
You	*stay*	*run*	+	You	*stay*	*run*
stay	2	0		*stay*	2	0
run	0	3		*run*	0	3

	b_a^1				ε_a^2	
Alice	*stay*	*run*	+	Alice	*stay*	*run*
stay	3	0		*stay*	3	0
run	0	2		*run*	0	2

utility for you is then the sum of these two utilities. For instance, your utility from choosing *stay* if your first-order expectation b_y^1 is *stay* and your second-order expectation ε_y^2 is $\frac{1}{2}(run + stay)$ is equal to $2 + \frac{1}{2}(2 + 0) = 3$. Similarly for Alice.

We now apply a simplified version of *iterated elimination of choices and first-order beliefs* to this game. As follows from yet another result in our companion paper Jagau and Perea (2017), a convenient way to capture the consecutive elimination steps in the present game goes by graphically depicting them as linear restrictions on the product space of first-order beliefs and second-order expectations for both you and Alice. To compute the set $R_y^1(1)$ of rational pairs of choices and first-order beliefs for you, we first depict the pairs (b_y^1, ε_y^2) of first-order beliefs and second-order expectations for which *stay* is rational, and the pairs for which *run* is rational. See the left-hand picture in Figure 1.

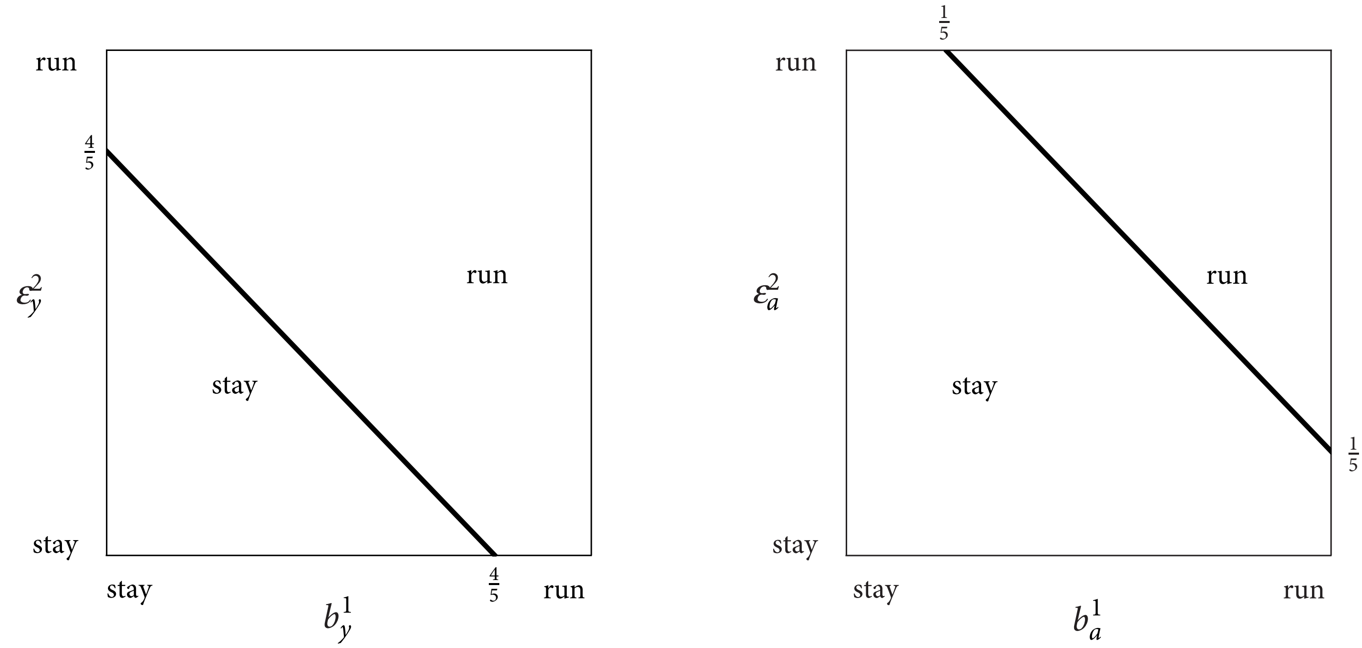
Note that *stay* can only be rational for a pair of beliefs and expectations (b_y^1, ε_y^2) if $b_y^1(run) \le \frac{4}{5}$. On the other hand, every first-order belief b_y^1 can be extended to a pair (b_y^1, ε_y^2) for which *run* is rational. Hence, we conclude that

$$R_y^1(1) = \left\{(stay, b_y^1) \middle| b_y^1(run) \le \frac{4}{5}\right\} \quad \cup \quad \{(run, b_y^1) | b_y^1 \in \Delta(\{stay, run\})\}.$$

In a similar way we can derive $R_a^1(1)$ from the right-hand picture of Figure 1 and conclude that

$$R_a^1(1) = \{(stay, b_a^1) | b_a^1 \in \Delta(\{stay, run\})\} \quad \cup \quad \left\{(run, b_a^1) \middle| b_a^1(run) \ge \frac{1}{5}\right\}.$$

Figure 1: Beliefs and Expectations for which Choices are Rational



The set of belief-expectation combinations $(b_y^1, \varepsilon_y^2))$ for which you believe in Alice's rationality is then given by the convex hull of $R_a^1(1)$. Graphically, this corresponds to the area above the thick line in the left-hand picture of Figure 2. Note that *stay* can only be rational for you for a pair of beliefs and expectations (b_y^1, ε_y^2) in $\mathrm{Conv}(R_a^1(1))$ if $b_y^1(run) \leq \frac{2}{3}$. On the other hand, every first-order belief b_y^1 can be extended to a pair (b_y^1, ε_y^2) in $\mathrm{Conv}(R_a^1(1))$ for which *run* is rational. Hence, we obtain that

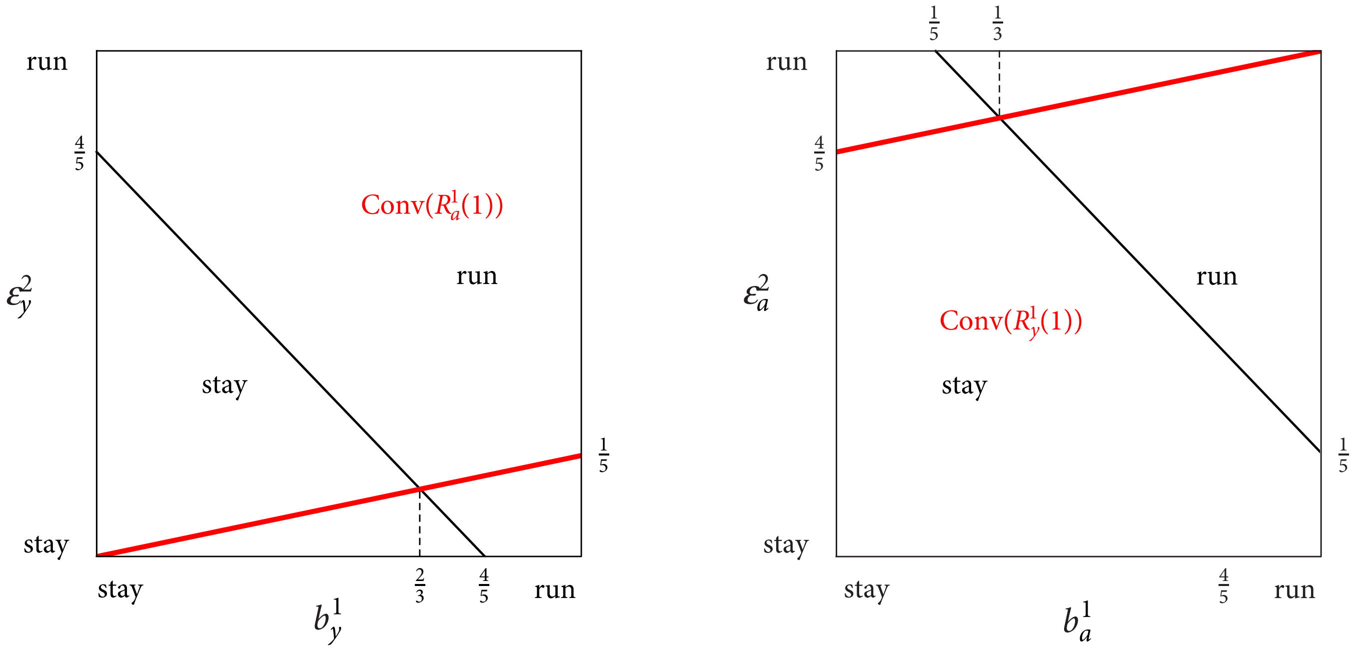
$$R_y^1(2) = \left\{(stay, b_y^1) \middle| b_y^1(run) \leq \frac{2}{3}\right\} \quad \cup \quad \{(run, b_y^1)|b_y^1 \in \Delta(\{stay, run\})\}.$$

Similarly, the convex hull of $R_y^1(1)$ is given by the area below the thick line in the right-hand picture of Figure 2. In the same way as above, we can derive from the right-hand picture of Figure 2 that

$$R_a^1(2) = \{(stay, b_a^1)|b_a^1 \in \Delta(\{stay, run\})\} \quad \cup \quad \left\{(run, b_a^1) \middle| b_a^1(run) \geq \frac{1}{3}\right\}.$$

If we were to continue in this fashion, we would see that $R_y^1(k) \neq R_y^1(k-1)$ and $R_a^1(k) \neq R_a^1(k-1)$ for every $k \geq 2$, and hence this procedure does not terminate within finitely many steps.

Figure 2: Convex hull of $R^1_a(1)$ and $R^1_y(1)$



Finally, it can be verified that

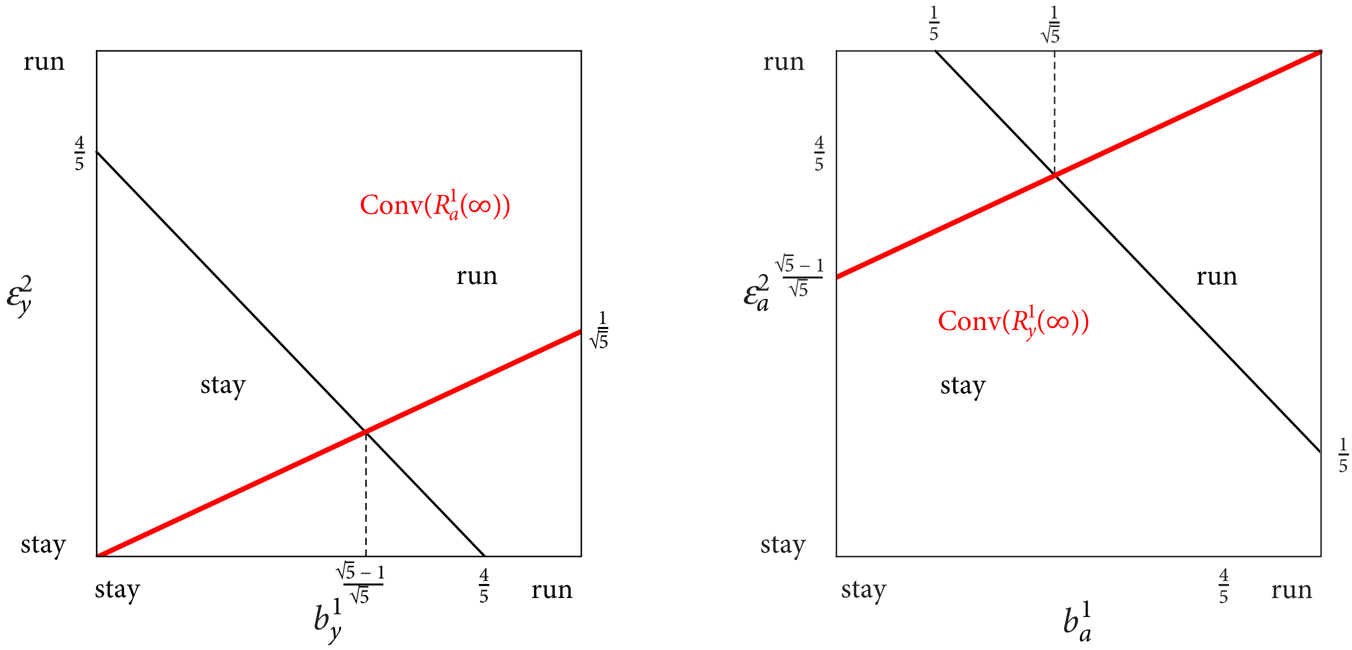
$$R^1_y(\infty) = \left\{ (stay, b^1_y) \middle| b^1_y(run) \leq \frac{\sqrt{5}-1}{\sqrt{5}} \right\} \quad \cup \quad \{(run, b^1_y) | b^1_y \in \Delta(\{stay, run\})\}$$

and

$$R^1_a(\infty) = \{(stay, b^1_a) | b^1_a \in \Delta(\{stay, run\})\} \quad \cup \quad \left\{ (run, b^1_a) \middle| b^1_a(run) \geq \frac{1}{\sqrt{5}} \right\},$$

where $\frac{\sqrt{5}-1}{\sqrt{5}} \approx 0.55$ and $\frac{1}{\sqrt{5}} \approx 0.45$. In particular, it follows that both you and Alice can rationally choose *stay* and *run* under common belief in rationality. Figure 3 shows how the sets $R^1_y(\infty)$ and $R^1_a(\infty)$ can be graphically constructed.

Figure 3: Convex hull of $R_a^1(\infty)$ and $R_y^1(\infty)$



VII Unilateral Games

None of the procedures studied in the previous two sections necessarily have a finite stopping time when the utilities of players depend on second-order beliefs or yet higher-order beliefs, different from what we are used to from traditional games. If utilities only depend on first-order beliefs and if the game is belief-continuous, such that common belief in rationality is characterized by iterated elimination of choices, we can find a bound on the number of steps that the procedure can possibly take: Given that the input for the procedure, i.e. $\times_{i\in I} C_i$ is finite, the number of elimination will be bounded at $\sum_{i\in I}(\#C_i - 1)$. If $n \geq 1$, such that utility depends on at least second-order beliefs, then the input for the elimination procedure becomes the uncountably infinite set $\times_{i\in I}(C_i \times B_i^n)$. Hence, even in a belief-continuous game, elimination of choices and nth-order beliefs need not converge after finitely many elimination steps if we do not make additional assumptions. In this section, we study a specific class of psychological games in which the utility function of one player depends on second-order beliefs and the utility functions of all other players depend only on first-order beliefs. For these games, elimination of choices and 1st-order beliefs can be shown to converge after finitely many steps.

Definition VII.1. *(Unilateral Psychological Game)*
A psychological game Γ is unilateral if the utility of one player depends only on second-order beliefs and all other players' utilities depend on first-order beliefs only.

Most examples of static psychological games that have been studied actually are unilateral psychological games (cf. Geanakoplos et al. 1989, Kolpin 1992).[10] Also, many dynamic psychological games can be associated with a static psychological game in this class as their normal form (examples are cf. Huang and Wu 1994, Dufwenberg 2002, Charness and Dufwenberg 2006, Battigalli and Dufwenberg 2007, Battigalli and Dufwenberg 2009).[11]

In what follows, we assume that player 1 cares about up to second-order beliefs and all other players care about first-order beliefs only.

We will now show that, for any unilateral game, iterated elimination of choices and 1st-order beliefs converges after finitely many steps and, more specifically, that the number of steps the procedure can take is bounded by $2(\#C_1 - 1) + \sum_{i\neq 1}(\#C_i - 1) + 1$.

Theorem VII.2. *(The Algorithm is Finite for Unilateral Games)*
For any unilateral game, iterated elimination of choices and 1st-order beliefs can take at most $2(\#C_1 - 1) + \sum_{i\neq 1}(\#C_i - 1) + 1$ elimination steps.

Proof.
Part 1 (Stopping Rule):
We start by showing that a universal stopping rule applies to iterated elimination of choices and 1st-order beliefs in all unilateral games:
Take a unilateral psychological game Γ and let there be a round $K \geq 0$ such that

$$\mathrm{proj}_{C_1} R_1^1(K) = \mathrm{proj}_{C_1} R_1^1(K+1) = \mathrm{proj}_{C_1} R_1^1(K+2)$$

and, for all players $i \neq 1$,

$$\mathrm{proj}_{C_i} R_i^1(K) = \mathrm{proj}_{C_i} R_i^1(K+1)$$

where $R_j^1(0) = C_j \times B_j^1$, $j \in I$.

Then $R_1^1(K+m) = R_1^1(K+2)$ and $R_i^1(K+m) = R_i^1(K+2)$, $i \neq 1$ for all $m \geq 2$. Hence, all choice-belief combinations in $R_j^1(K+2)$, $j \in I$, are consistent with common belief in rationality.

[10] A notable exception is Rabin's (1993) model of reciprocity in which all players are allowed to care about second-order beliefs.

[11] Not every dynamic psychological game can necessarily be associated with a static psychological game in this manner. If players' utility functions for some histories of play depend on updated beliefs, it is unclear how a sensible normal-form version of the corresponding dynamic game would look like. Examples of such irreducible dynamic psychological games include the dynamic prisoner's dilemma from Geanakoplos et al. (1989) and the reciprocity model of Dufwenberg and Kirchsteiger (2004). It seems an interesting avenue for future work to extend our characterizations to modes of reasoning in dynamic psychological games. Falk and Fischbacher (2006) study a model of sequential reciprocity in which players care about initial first- and second-order beliefs. So each game studied by these authors has a corresponding normal-form psychological game, but not all of these are necessarily unilateral.

To prove the stopping rule, first note that, for any $k \geq 1$, and for all players $i \neq 1$ who only care about first-order beliefs, we can write

$$
\begin{aligned}
R_i^1(k) = & \{(c_i, b_i^1) \in R_i^1(k-1) | \exists b_i^2 \in \Delta(R_{-i}^1(k-1)) \text{ with } \operatorname{marg}_{X_i^1} b_i^2 = b_i^1 \\
& \text{such that } u_i(c_i, b_i^1) \geq u_i(c_i', b_i^1), \forall c_i' \in C_i\} \\
= & \{(c_i, b_i^1) \in R_i^1(k-1) | b_i^1 \in \Delta(\operatorname{proj}_{C_{-i}} R_{-i}^1(k-1)) \text{ and } u_i(c_i, b_i^1) \geq u_i(c_i', b_i^1), \forall c_i' \in C_i\}.
\end{aligned}
$$

Given this simplification, we can now easily see that $\operatorname{proj}_{C_1} R_1^1(K) = \operatorname{proj}_{C_1} R_1^1(K+1)$ and $\operatorname{proj}_{C_i} R_i^1(K) = \operatorname{proj}_{C_i} R_i^1(K+1)$, $i \neq 1$ imply that $R_i^1(K+1) = R_i^1(K+2)$, $i \neq 1$. By definition of $R_1^1(K+3)$ it immediately follows that $R_1^1(K+2) = R_1^1(K+3)$.

Now since clearly $\operatorname{proj}_{C_i} R_i^1(K+1) = \operatorname{proj}_{C_i} R_i^1(K+2)$, $i \neq 1$ and, by assumption, $\operatorname{proj}_{C_1} R_1^1(K+1) = \operatorname{proj}_{C_1} R_1^1(K+2)$, we can also conclude that $R_i^1(K+2) = R_i^1(K+3)$, $i \neq 1$.

But then all reductions have already converged so

$$R_j^1(K+m) = R_j^1(K+2)$$

for all players j as desired.

Part 2 (Upper Bound):

As the stopping rule shows, not eliminating any choices at a step $K+1$ of the algorithm for any player is already enough to conclude that the reductions of all players $i \neq 1$ do not change in the next step $K+2$. So if the algorithm does not eliminate choices at step $K+1$, it must already converge *unless* choices for player 1 get eliminated at round $K+2$. So before convergence, there can at most be gaps of one round where the algorithm eliminates no choices and there can be at most as many such gaps as we can eliminate choices for player 1. This way, we can conclude that elimination of choices and 1st-order beliefs can take at most $2(\#C_1 - 1) + \sum_{i \neq 1}(\#C_i - 1)$ steps before the last choice gets eliminated. Noting that from eliminating the last choice it takes another step until the reduced sets of first-order beliefs converge we receive the desired upper bound. □

If a unilateral game is not belief-continuous, *iterated elimination of choices and first-order beliefs* still converges after a maximum of $2(\#C_1 - 1) + \sum_{i \neq 1}(\#C_i - 1) + 1$ steps, but it need not necessarily yield exactly the choice-belief combinations that are consistent with common belief in rationality. In addition, it can select choice-belief combinations that get eliminated at the limit. To tightly characterize common belief in rationality here, we would need to use *iterated elimination of choices and second- and higher-order beliefs* (cf. procedure VI.4) which is not in general a finite procedure.

At least for some unilateral games that bound is now tight, as follows from observation VII.3 below. In the observation, we present a class of 2-player games in which player 1 has n choices, player 2 has 2 choices, and the algorithm converges after $2n$ steps. What can be nicely seen in this class of games is that, different from traditional games, we can consecutively eliminate all but one choice for player 1 without any intermediate elimination of choices for opponents. For unilateral games, this pattern can occur as long as at least one opponent has a non-singleton choice set. By contrast, in a traditional game, we can never eliminate choices for one player in several consecutive elimination steps without eliminating choices for other players as well.

Observation VII.3. *(The bound is tight for $n \times 2$-games)*
For any n, there exists an $n \times 2$-unilateral game such that the algorithm converges after exactly $2n$ elimination steps.

Proof.
We construct an $n \times 2$ unilateral game with the desired properties. Let $C_1 = \{a_1, \dots, a_n\}$ and $C_2 = \{b, c\}$. Define player 1's utility as follows

$$u_1(a_1, b_1^2) = 1 \text{ for all } b_1^2 \in B_1^2.$$

$$u_1(a_2, b_1^2) = \begin{cases} 1, & b_1^2 = (b, a_1) \\ 2, & \text{else} \end{cases}$$

$$u_1(a_3, b_1^2) = \begin{cases} 1, & b_1^2 = (b, a_1) \\ 2, & b_1^2 = (c, a_1) \\ 3, & \text{else} \end{cases}$$

$$u_1(a_4, b_1^2) = \begin{cases} 1, & b_1^2 = (b, a_1) \\ 2, & b_1^2 = (c, a_1) \\ 3, & b_1^2 = (c, a_2) \\ 4, & \text{else} \end{cases}$$

$$\vdots$$

$$u_1(a_n, b_1^2) = \begin{cases} 1, & b_1^2 = (b, a_1) \\ 2, & b_1^2 = (c, a_1) \\ 3, & b_1^2 = (c, a_2) \\ \vdots & \\ n-1, & b_1^2 = (c, a_{n-2}) \\ n, & \text{else} \end{cases}$$

And let player 2's utility be defined by

$$u_2(b, b_2^1) = b_2^1(a_n) \text{ and } u_2(c, b_2^1) = b_2^1(a_1)$$

for all $b_2^1 \in B_2^1$ We now perform iterated elimination of choices and first-order beliefs on this game:

- *Step 1:* $R_1^1(1) = \{(a_1, b_1^1) | b_1^1(b) = 1\} \cup \bigcup_{i=2}^{n-1} \{(a_i, b_1^1) | b_1^1(b) = 1 \text{ or } b_1^1(c) = 1\} \cup \{(a_n, b_1^1) | b_1^1 \in B_1^1\}$,
  $R_2^1(1) = \{(b, b_2^1) | b_2^1(a_n) \geq b_2^1(a_1)\} \cup \{(c, b_2^1) | b_2^1(a_1) \geq b_2^1(a_n)\}$
- *Step 2:* $R_1^1(2) = \bigcup_{i=2}^{n-1} \{(a_i, b_1^1) | b_1^1(c) = 1\} \cup \{(a_n, b_1^1) | b_1^1 \in B_1^1\}$, $R_2^1(2) = R_2^1(1)$
- *Step 3:* $R_1^1(3) = R_1^1(2)$, $R_2^1(3) = \{(b, b_2^1) | b_2^1(a_1) = 0\} \cup \{(c, b_2^1) | b_2^1(a_1) = b_2^1(a_n) = 0\}$
- *Step 4:* $R_1^1(4) = \bigcup_{i=3}^{n-1} \{(a_i, b_1^1) | b_1^1(c) = 1\} \cup \{(a_n, b_1^1) | b_1^1 \in B_1^1\}$, $R_2^1(4) = R_2^1(3)$
- *Step 5:* $R_1^1(5) = R_1^1(4)$,
  $R_2^1(5) = \{(b, b_2^1) | b_2^1(a_1) = b_2^1(a_2) = 0\} \cup \{(c, b_2^1) | b_2^1(a_1) = b_2^1(a_2) = b_2^1(a_n) = 0\}$
- *Step 6:* $R_1^1(4) = \bigcup_{i=4}^{n-1} \{(a_i, b_1^1) | b_1^1(c) = 1\} \cup \{(a_n, b_1^1) | b_1^1 \in B_1^1\}$, $R_2^1(6) = R_2^1(5)$

$\vdots$

- *Step* $2n-4$*:* $R_1^1(2n-4) = \{(a_{n-1}, b_1^1) | b_1^1(c) = 1\} \cup \{(a_n, b_1^1) | b_1^1 \in B_1^1\}$,
  $R_2^1(2n-4) = R_2^1(2n-5) =$
  $\{(b, b_2^1) | b_2^1(a_1) = \cdots = b_2^1(a_{n-3}) = 0\} \cup \{(c, b_2^1) | b_2^1(a_1) = \cdots = b_2^1(a_{n-3}) = b_2^1(a_n) = 0\}$
- *Step* $2n-3$*:* $R_1^1(2n-3) = R_1^1(2n-4)$,
  $R_2^1(2n-3) =$
  $\{(b, b_2^1) | b_2^1(a_1) = \cdots = b_2^1(a_{n-2}) = 0\} \cup \{(c, b_2^1) | b_2^1(a_1) = \cdots = b_2^1(a_{n-2}) = b_2^1(a_n) = 0\}$
- *Step* $2n-2$*:* $R_1^1(2n-2) = \{(a_n, b_1^1) | b_1^1 \in B_1^1\}$, $R_2^1(2n-2) = R_2^1(2n-3)$
- *Step* $2n-1$*:* $R_1^1(2n-1) = R_1^1(2n-2)$, $R_2^1(2n-1) = \{(b, b_2^1) | b_2^1(a_n) = 1\}$
- *Step* $2n$*:* $R_1^1(2n) = \{(a_n, b_1^1) | b_1^1(b) = 1\}$, $R_2^1(2n) = R_2^1(2n-1)$

$\square$

VIII Discussion

A. Alternative Modeling Approaches

A reader familiar with the existing psychological-games literature will likely have noticed that in definition II.2 we model static psychological games slightly differently than previous contributions. Here, we show that the different modeling approaches are entirely equivalent.

The best-known modeling approaches in the literature are the ones from Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009). We can easily convince ourselves that our definition of static psychological games captures no more and no less than the alternative definitions used in these papers:

Geanakoplos et al.'s (1989) Approach:
In Geanakoplos et al. (1989), players' utility is defined to be a function

$$u_i : \bigtimes_{j \in I} C_j \times B_i \to \mathbb{R}.$$

Players then choose $\sigma_i \in \Delta(C_i)$ to maximize the expected value

$$\overline{u}_i(\sigma_i, \sigma_{-i}, b_i) = \sum_{c_i \in C_i} \sum_{c_{-i} \in C_{-i}} \sigma_i(c_i)\sigma_{-i}(c_{-i})u_i(c_i, c_{-i}, b_i).$$

where $\sigma_{-i} = (\sigma_j)_{j \neq i}$ is a vector of opponents' randomized choices.

The authors interpret $\overline{u}_i$ to be the payoff of player i if he "believed b_i and then found out that σ was actually played". So the distribution generating σ captures objective probabilities and b_i subjective ones. Still, players maximize $\overline{u}_i$. So in some way, they know the distribution σ, though they are "presumed not to observe the mixture". Keeping up the distinction between objective σ_{-i} and subjective b_i^1 leads to obvious inconsistencies. In the case $\sigma_{-i} \neq b_i^1$, player i believes that opponents choose according to the distribution b_i^1 while, at the same time, maximizing under the assumption that they choose according to σ_{-i}. Allowing for this configuration does not seem to be particularly useful. Probably this does not hamper the analysis in that paper because all results of Geanakoplos et al. (1989) are derived under a correct beliefs assumption so that, automatically, $\sigma_{-i} = b_i^1$. In what follows, we will therefore also assume $b_i^1 = \sigma_{-i}$. Then $\overline{u}_i$ becomes

$$\sum_{c_i \in C_i} \sigma_i(c_i) \sum_{c_{-i} \in C_{-i}} b_i^1(c_{-i})u_i(c_i, c_{-i}, b_i).$$

Clearly, defining a value function $\hat{u}_i(c_i, b_i) = \sum_{c_{-i} \in C_{-i}} b_i^1(c_{-i})u_i(c_i, c_{-i}, b_i)$ maps this directly into our framework.

Hence, up to allowing players to select randomized choices σ_i, this modeling approach is entirely equivalent to the one we take in definition II.2.

Battigalli and Dufwenberg's (2009) Approach:

In Battigalli and Dufwenberg (2009), a player's utility in a dynamic game is defined to be

$$u_i : Z \times \bigtimes_{j \in I} B_j \times S_{-i} \to \mathbb{R}$$

where Z are the "terminal nodes" of the game and S_{-i} are opponents' strategies.

In a static game, a special case of Battigalli and Dufwenberg's (2009) dynamic framework, Z can be identified with $\bigtimes_{j \in I} C_j$, since every combination of players' choices can be identified with a unique "history". Also $C_{-i} = S_{-i}$ such that the dependence on opponents' strategies becomes redundant.

Thus, Battigalli and Dufwenberg's (2009) choice of utility for static psychological games can be written as

$$u_i : \bigtimes_{j \in I} (C_j \times B_j) \to \mathbb{R}.$$

So their utility function generalizes Geanakoplos et al.'s (1989) approach by allowing for functional dependence on opponents' belief hierarchies $\bigtimes_{j \neq i} B_j$.

Players are then assumed to choose $c_i \in C_i$ to maximize

$$E_{b_i}[u_i(c_i, c_{-i}, b_i, b_{-i})] = \int_{B_{-i}} \left(\sum_{c_{-i} \in C_{-i}} b_i^1(c_{-i}) u_i(c_i, c_{-i}, b_i, b_{-i}) \right) \mathrm{d}b_i,$$

where b_i is identified with a probability measure over $\Delta(C_{-i} \times B_{-i})$. Defining the value function

$$\hat{u}_i(c_i, b_i) = E_{b_i}[u_i(c_i, c_{-i}, b_i, b_{-i})]$$

we see that also this modeling approach can be mapped into our framework and vice versa.

In general, our approach can be said to differ from the ones discussed here in that it assumes a *one-person perspective* relative to a psychological game: When making a choice c_i, a given player i forms beliefs over choices of opponents, first-order beliefs of opponents, second-order beliefs of opponents, and so on; and all these beliefs and their interrelations are encapsulated in the belief hierarchy b_i. By defining utility over choices and belief hierarchies, we distinguish between the variable a player can influence (c_i) and every decision-relevant information that he cannot influence (b_i). Since every piece of information encoded in b_i can become utility relevant in a psychological game, it seems natural to make no further distinctions between the beliefs $b_i^1, b_i^2, \ldots$ when defining utility.

Both Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009) assume an *observers' perspective*. In Battigalli and Dufwenberg (2009), players directly care about their own and opponents' choice-belief combinations $(c_i, b_i)_{i \in I}$ and when making a choice c_i, a given player i then forms an expectation regarding the (to him) unknown parameters (c_{-i}, b_{-i}) where this expectation follows from his belief hierarchy b_i when interpreted as a probability measure on $C_{-i} \times B_{-i}$. Clearly, this is ultimately no different from directly starting off with utilities defined on $C_i \times B_i$. What is *conceptually* different here is that objects that are "in a players' head" (c_i and b_i) get distinguished from objects that are "out there" (c_{-i} and b_{-i}). While it might seems philosophically attractive to make this distinction, it is important to note that we can never really keep the two types of objects separated. All information regarding c_{-i} and b_{-i} that is used by a player i is already present in b_i in a world of coherent beliefs and i can never consistently do any different than looking into his own belief hierarchy when forming expectations about these things. Also, it is not really clear where we would stop the separation. Clearly, first-order beliefs b_i^1 should be identified with beliefs about opponents' choices c_{-i}. But also second order beliefs b_i^2 should be identified with beliefs about opponents' choice-first-order-belief combinations (c_{-i}, b_{-i}^1). Clearly, we can indefinitely proceed this way, replacing things in "i's head" with things that are "out there", since there is just no distinction made between these categories in the setup of a psychological game. We therefore opt to refrain from assuming any such distinctions into the model and thereby keep the definition of utility functions as parsimonious as possible. Clearly, which approach is preferred is ultimately a purely conceptual question that does not matter for the analysis carried out here or in Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009).

B. Summary and Conclusion

Since its introduction by Geanakoplos et al. (1989), psychological game theory has become increasingly popular in applications as a tool to capture numerous belief-dependent motivations and emotional mechanisms in a natural way. Nevertheless, our theoretical understanding of psychological games still falls far short of what we would be used to from traditional games.

In this paper we started theorizing at square one and provided a systematic analysis of common belief in rationality in static psychological games. While we restricted our analysis to static psychological games in the sense of Geanakoplos et al. (1989) and thereby excluded the rich classes of dynamic psychological games in which utility is allowed to depend on updated beliefs (cf. Battigalli and Dufwenberg 2009), we expect our core results to carry over to this richer class after appropriate modifications.[12] Also, it seems clear that the computational issues raised by us will also arise in dynamic psychological games which are typically even harder to analyze than static psychological games.

Our results not only relax the previously known existence conditions for common belief in rationality in psychological games, but also provide iterative procedures that select the choices that can be made under common belief in rationality in a given psychological game. As we saw, special classes of psychological games allow for massive simplifications in these algorithms relative to the more general case.

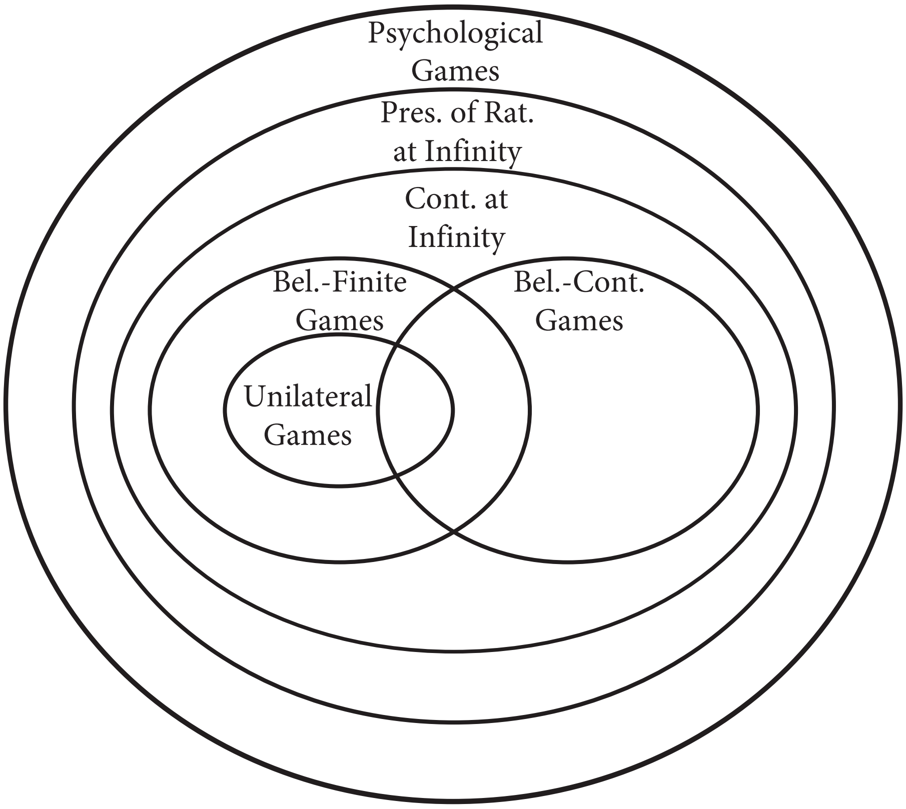
[12] Part of this extension is already provided in the aforementioned master thesis by Sanna (2016).

Together with the different existence results, we receive an extensive classification of psychological games that is summarized in table 6 below. We accompany the table by a set diagram (figure 4) that illustrates how different classes of psychological games that appeared in this paper relate to each other.

Table 6: Psychological Games and their Properties

	Possibility of Common Belief in Rationality	Existence of Psychological Nash Equilibrium	Characterizing Algorithm for Common Belief in Rationality
All Games (Definition II.2)	Not guaranteed (example IV.3)	Not guaranteed (example IV.13)	Elimination of choices and belief hierarchies (procedure V.2)
Preservation of Rationality at Infinity (Definition IV.1)	Guaranteed (theorem IV.2)	⋮	⋮
Continuity at Infinity (Definition IV.4)	⋮	⋮	⋮
Belief-Continuous Games (Definition IV.9)	⋮	Guaranteed (theorem IV.12)	⋮
Belief-Finite Games (Definition IV.6)	⋮	Not guaranteed (example IV.13)	Elimination of choices and nth- and higher-order beliefs (procedure VI.4)
Belief-Finite, Belief-Continuous Games	⋮	Guaranteed (theorem IV.12)	Elimination of choices and nth-order beliefs (procedure VI.1)
Belief Continuous, Unilateral Games (Definitions IV.9,VII.1)	⋮	⋮	Elimination of choices and 1st-order beliefs (finite by theorem VII.2)

Figure 4: Classes of Psychological Games



Bibliography

Akerlof, R., 2017: Value formation: The role of esteem. *Games and Economic Behavior*, **102 (1)**, 1–19.

Attanasi, G., P. Battigalli, and E. Manzoni, 2016: Incomplete-information models of guilt aversion in the trust game. *Management Science*, **62 (3)**, 648–667.

Attanasi, G., P. Battigalli, and R. Nagel, 2017: Disclosure of belief-dependent preferences in a trust game, working paper.

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

Battigalli, P., G. Charness, and M. Dufwenberg, 2013: Deception: The role of guilt. *Journal of Economic Behavior & Organization*, **93**, 227–232.

Battigalli, P., and M. Dufwenberg, 2007: Guilt in games. *American Economic Review*, **97 (2)**, 170–176.

Battigalli, P., and M. Dufwenberg, 2009: Dynamic psychological games. *Journal of Economic Theory*, **144 (1)**, 1–35.

Battigalli, P., M. Dufwenberg, and A. Smith, 2015: Frustration and anger in games, working paper.

Battigalli, P., and M. Siniscalchi, 2002: Strong belief and forward induction reasoning. *Journal of Economic Theory*, **106 (2)**, 356–391.

Bjorndahl, A., J. Y. Halpern, and R. Pass, 2016: Language-based games, working paper.

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

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

Caplin, A., and J. Leahy, 2004: The supply of information by a concerned expert. *Economic Journal*, **114 (497)**, 487–505.

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

Dufwenberg, M., 2002: Marital investments, time consistency and emotions. *Journal of Economic Behavior & Organization*, **48 (1)**, 57–69.

Dufwenberg, M., and G. Kirchsteiger, 2004: A theory of sequential reciprocity. *Games and Economic Behavior*, **47 (2)**, 268–298.

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

Dufwenberg, M., Jr., and M. Dufwenberg, 2016: Lies in disguise a theoretical analysis of cheating, working paper.

Falk, A., and U. Fischbacher, 2006: A theory of reciprocity. *Games and Economic Behavior*, **54 (2)**, 293–315.

Geanakoplos, J., D. Pearce, and E. Stacchetti, 1989: Psychological games and sequential rationality. *Games and Economic Behavior*, **1 (1)**, 60–79.

Huang, P. H., and H.-M. Wu, 1994: More order without more law: A theory of social norms and organizational cultures. *Journal of Law, Economics, & Organization*, **10 (2)**, 390–406.

Huck, S., and D. Kübler, 2000: Social pressure, uncertainty, and cooperation. *Economics of Governance*, **1**, 199–212.

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

Khalmetski, K., A. Ockenfels, and P. Werner, 2015: Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, **159**, 163–208.

Kolpin, V., 1992: Equilibrium refinement in psychological games. *Games and Economic Behavior*, **4 (2)**, 218–231.

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

Li, J., 2008: The power of conventions: A theory of social preferences. *Journal of Economic Behavior & Organization*, **65 (3)**, 489–505.

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

Nozick, R., 1969: Newcomb's problem and two principles of choice. *Essays in Honor of Carl G. Hempel*, N. Rescher, Ed., Berlin/Heidelberg: Springer.

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

Perea, A., 2012: *Epistemic Game Theory: Reasoning and Choice.* Cambridge: Cambridge University Press.

Rabin, M., 1993: Incorporating fairness into game theory and economics. *American Economic Review*, **83 (5)**, 1281–1302.

Sanna, F. A., 2016: Universal spaces of hierarchies of beliefs: an application to k-th order psychological games, unpublished master thesis, supervised by P. Battigalli and S. C. Vioglio.

Sebald, A., 2010: Attribution and reciprocity. *Games and Economic Behavior*, **68 (1)**, 339–352.

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