Common Belief in Rationality in Psychological Games

Stephan Jagau[ABC] and Andrés Perea[BD]

February 2, 2018

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 behavior (Dufwenberg and Dufwenberg 2016), surprise (Khalmetski et al. 2015), and anger (Battigalli et al. 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 existing contributions 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 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 that translates Geanakoplos et al.'s (1989) continuity condition to their more general framework.

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. Static psychological games as defined by Geanakoplos et al. (1989) differ from dynamic psychological games both in that players are allowed to move sequentially and in that their preferences may depend on updated beliefs that arise during the play of the dynamic game. While our results here are restricted to static psychological games, this restriction is made for clarity of exposition and not because we believe our results do not extend to dynamic psychological games as defined in Battigalli and Dufwenberg (2009). In dynamic games, common belief in rationality does not restrict the way players update their beliefs as the game unfolds – different from stronger reasoning concepts such as common belief in future rationality (Dekel et al. 1999, 2002, Asheim and Perea 2005, Perea 2014) and common strong belief in rationality (Pearce 1984, Battigalli 1997, Battigalli and Siniscalchi 2002). Therefore, allowing for dynamic games and dependence of preferences on updated beliefs in our investigation would only lead to a more complex and less accessible notational apparatus, but not to qualitatively different results.[1] Since the restriction to static psychological games is made primarily for pedagogical reasons, we will use the terms "psychological game" and "static psychological game" interchangeably in the remainder of the paper and make differences explicit where they matter.

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. (2013). 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 (2013) CR-condition to a broader class of psychological games.

[1] A sketch of how definitions and results would carry over to dynamic games is available from the authors upon request. Extending our investigation to common belief in future rationality and common strong belief in rationality in dynamic psychological games is an interesting avenue for future research.

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 this paper, in particular, we provide conditions under which the applicable algorithm is of finite length like iterated elimination of strictly dominated choices is in traditional games.

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. 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 *linearly* care about *expected values* of finite levels of higher-order beliefs (cf. e.g. Rabin 1993, Dufwenberg and Kirchsteiger 2004, Battigalli and Dufwenberg 2007, Battigalli et al. 2017). In our companion paper Jagau and Perea (2017), we provide a first formal definition of this class of psychological games. We show that these games 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.

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 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 show that the appropriate algorithm *iterated elimination of choices and 1st-order beliefs* is of finite length. 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

We start by giving a formal definition of static psychological games:

Definition II.1. *(Static Psychological Game)*
A **static psychological game** *is a tuple* $\Gamma = (C_i, B_i, u_i)_{i \in I}$ *with* I *a finite set of players,* C_i *the finite set of choices available to player* i*,* B_i *the set of belief hierarchies for player* i *expressing coherency and common belief in coherency and* u_i *a utility function of the form*

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

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.

Each belief hierarchy b_i is a chain of probability distributions $(b_i^1, b_i^2, \dots)$ 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 B_i^n, $n \geq 1$ of nth-order beliefs and the set B_i of belief hierarchies expressing coherency and common belief in coherency can be recursively constructed. In appendix A we redo their construction for our specific setup.

Here, we only 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.

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

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 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.

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*. To start, we define this condition formally and show by means of a constructive proof that it ensures the existence of belief hierarchies that express common belief in rationality. Subsequently, 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. We then compare our condition to previous existence conditions, showing that preservation of rationality at infinity significantly expands the scope of games for which the possibility of common belief in rationality is ensured.

IV.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.[2]

Definition IV.1. *(Preservation of Rationality at Infinity)*
Let $\Gamma = (C_i, B_i, u_i)_{i \in I}$ be a psychological game and let $c_i \in C_i$ be a choice and $b_i \in B_i$ a belief hierarchy for some player $i \in I$ in it. Suppose that, for every $n \geq 1$, there is some $\hat{b}_i \in B_i$ with $\hat{b}_i^n = b_i^n$ such that c_i is rational for $\hat{b}_i$. The game is said to preserve rationality at infinity if choice c_i is then also rational for 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. 2013) 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 $b_i^n[c^1, c^2, ...] = (c^1, ..., c^n)$.

[2] A stronger existence condition, *continuity at infinity*, would require that utilities satisfy

$$\lim_{n \to \infty} \left(\sup_{b_i, \hat{b}_i :\, b_i^n = \hat{b}_i^n} |u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| \right) = 0$$

for all $c_i \in C_i$ and every player i. In words, utility may depend on all levels of higher-order beliefs, but the impact that a specific level n has on the overall utility vanishes as n becomes large. This condition would be intermediate between preservation of rationality at infinity and belief continuity (see section IV.B) and it directly translates to *continuity at infinity* as defined for infinitely repeated traditional games (see Fudenberg and Levine 1983) to our setup. Conversely, it is straightforward to define an easy-to-check repeated-games counterpart of preservation of rationality at infinity. It would be interesting to investigate whether existence conditions in that literature can still be weakened by replacing continuity at infinity with that condition.

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 $d(1) := (d_i(1))_{i \in I}$. Then, for all players i, define a new belief hierarchy

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

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. That is, we must show, for every $n \geq 1$ and every player j, that choice e_j^n is rational for the belief hierarchy $b_j[e^{n+1}, e^{n+2}, ...]$. Fix such an n and a player j.

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

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

where $g^{n+m+1}, g^{n+m+2}, ...$ are choice-combinations in $\times_{i \in I} C_i$.

Define the belief hierarchy

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

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

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

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:[3]

[3] An example with the same structure, the *deeply surprising proposal*, has independently been developed by Bjorndahl et al. (2013).

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 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 $b_1^{timid} = b_1[(timid, *), (timid, *), \dots]$. 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.

As example IV.3 and our theorem IV.2 show, common belief in rationality can only ever fail in psychological games where utility exhibits a peculiar type of discontinuous dependence on the full belief hierarchy of players. Under these preconditions, it seems highly unlikely that we would ever run into this problem in real-life applications of psychological games. Still, one might ask how useful our existence condition *preservation of rationality at infinity* is relative to what we already knew about the possibility of common belief in rationality in psychological games from Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009). In the remainder of this section, we therefore zoom in on the relation between our existence condition and theirs. Firstly, we show that the previous existence conditions are implied by our condition in the class of games we consider. Secondly, we provide a simple example of a game for which our condition guarantees existence while the previously known ones do not.

IV.B 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.*[4]

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 *preservation of rationality at infinity.* Also, it is easy to come up with examples where our existence condition reaches beyond belief continuity.

We start by defining belief continuity as introduced in Geanakoplos et al. (1989). To formally define this property, let $d(b_i^k, \hat{b}_i^k)$ 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 $\hat{d}(b_i, \hat{b}_i) = \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k d(b_i^k, \hat{b}_i^k)$. It is well known that the distance $\hat{d}$ then metricizes the product space B_i. Given these preliminaries, we define:

Definition IV.4. *(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* $\hat{d}(b_i, \hat{b}_i) < \delta$ *we have that* $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$*.*

Coming from this original definition of belief continuity, a more intuitive and easy-to-check characterization goes in terms of trembles of finite levels of higher-order beliefs:

Lemma IV.5. *(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 hierarchy* $\hat{b}_i$ *with* $d(b_i^m, \hat{b}_i^m) < \delta$ *for all* $m \leq k$ *we have that* $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$*.*

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 $\hat{d}(b_i, \hat{b}_i) < \delta$.

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

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

[4]For dynamic games, Battigalli and Dufwenberg (2009) study common strong belief in rationality. So their existence result goes even a little further 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.

where for the first inequality we used $d(b_i^m, \hat{b}_i^m) \leq 1$ 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 $d(b_i^m, \hat{b}_i^m) < \delta$ for all $m \leq k$ implies $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$.

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

Then

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

So, in particular,

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

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

□

Going from here, it is straightforward to show that *belief continuity* implies our condition *preservation of rationality at infinity*:

Theorem IV.6. *(Belief Continuity Refines Preservation of Rationality at Infinity)*
If a game is belief-continuous, then it preserves rationality at infinity.

Proof. Consider a game $\Gamma = (C_i, B_i, u_i)_{i \in I}$ that is belief-continuous, 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 belief-continuous there are, by Lemma IV.5, an $n \geq 1$ and a δ 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 $d(\hat{b}_i^m, b_i^m) < \delta$, $m \leq n$. In particular, it follows that

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

since, by definition, $d(\hat{b}_i^m, b_i^m) = 0$, $m \leq n$. 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. □

It is now clear that our existence condition *preservation of rationality at infinity* nests all previously known existence results for static psychological games.

As shown in Geanakoplos et al. (1989), requiring belief continuity even ensures a little more than just the possibility of common belief in rationality. On top of this, there exists a *psychological Nash equilibrium* in every belief-continuous game. That is, we can find a combination of *simple belief hierarchies* for all players that expresses common belief in rationality.[5]:

Definition IV.7. *(Psychological Nash Equilibrium)*
Let $\sigma \in \times_{i \in I} \Delta(C_i)$ be a vector of probability distributions over players' choices and 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. σ constitutes 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 $c_i' \in C_i$.*

Theorem IV.8. *(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. □

To show where *preservation of rationality at infinity* does reach beyond *belief continuity*, we provide a slightly modified version of example IV.3:

Example IV.9. (Common Belief in Rationality without Belief Continuity)
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 $B_1(*, timid)$ be 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.

[5] For an in-depth treatment for traditional games see Perea 2012.

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

It is easy to see that this game preserves rationality at infinity: Since utilities depend on at most second-order beliefs, c_i is necessarily rational for b_i whenever c_i is rational for some $\hat{b}_i$ with $\hat{b}_i^2 = b_i^2$. However, the game is not belief-continuous since slightly perturbing second-order beliefs for any $b_1 \in B_1(*, timid)$ leads to discontinuous changes in $u_1(bold, b_1)$.

So while *belief-continuity* does not allow us to ascertain the possibility of common belief in rationality, *preservation of rationality at infinity* does. In fact, it is easy to find belief hierarchies that rationalize either choice for player 1 while expressing common belief in rationality: To do this, we vary the technique 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,

$$\begin{aligned} b_i(1) &= b_i[d(1), c^1, c^2, \dots] \\ b_i(2) &= b_i[d(2), d(1), c^1, \dots] \\ b_i(k) &= b_i[d(k), d(k-1), d(k-2), \dots] \end{aligned}$$

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 hierarchy $\hat{b}_i = b_i[(bold, *), (bold, *), (timid, *), (timid, *), (bold, *), (bold, *), \dots]$ is 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.

While the Modified Bravery Game II *must* allow for common belief in rationality by our theorem IV.2, the fact that utilities are not belief-continuous leaves it open whether a psychological Nash equilibrium exists. To conclude, we show that there is indeed no equilibrium.

To see this, note that there is no *simple* belief hierarchy that expresses common belief in rationality in this game: Player 1 *strictly* prefers to choose *timid* whenever $b_1 \notin B_1(*, timid)$ and *strictly* prefers to choose *bold* otherwise. So the only candidates for his equilibrium belief hierarchy would be the two deterministic belief hierarchies $b_1[(timid, *), (timid, *), \dots]$ 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.

The game from example IV.9 might appear quite artificial, but it encapsulates a highly-relevant psychological phenomenon: In many experimental and real-life risky decisions, people are prone to the *certainty effect* (Tversky and Kahneman 1981, 1986). Moving from almost certainty to certainty of an event can discontinuously change the evaluation of alternatives and thereby dramatically change behavior. Given the prevalence of the certainty effect in individual-decision settings, it is plausible that similar discontinuities can also play a role when agents reason about others' intentions and beliefs. Clearly, whenever we want to model a game in ways that take account of the certainty effect and similar discontinuities in the processing of subjective probabilities, we will automatically venture outside the class of *belief-continuous* games. At the same time, already the fact that people in real-life decision problems plausibly care about at most finite levels of higher-order beliefs puts us squarely within the realm of games that *preserve rationality at infinity*.

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, we start by presenting 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.

V.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 consider ditching her in the first place. In particular, you get more enjoyment out of ditching Alice the more you think she expects that you go to the bar. If you ditch her, it is clear that there will not be another date. So given that you decide to ditch Alice, you do not care whether she comes to the bar or not.

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*.

Since b_a^1, b_y^1, and ε_y^2 are all probabilities and since the utility functions of you and Alice are *linear* in those probabilities, we can conveniently depict utilities by finite matrices as we are used to do in traditional static games. In particular, 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 3 below.[6]

[6] This depiction instances a more general representation theorem for so-called *additive expectation-based* games in our companion paper Jagau and Perea (2017).

Table 3: 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

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 your choice *date*. Because your utility is linear in this second-order expectation, we can summarize that component of u_y by collecting 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.[7] 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.

[7] 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 a *unilateral, additive 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.

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. 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.

V.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.[8] One might expect that things get simpler in interesting special cases of psychological games. In sections VI and VII we will discuss two such cases.

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:

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

[8] 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.

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

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 $\text{proj}_{C_i}(R_i(k+1))$.

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 $\text{proj}_{C_i}(R_i(\infty))$.

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 belief hierarchies 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 $\Delta(R_{-i}(k))$.

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.

That algorithm V.2 does characterize common belief in rationality in static psychological games might hardly seem surprising. In essence, it is a restatement of definition III.3 as an algorithm. The more interesting question is whether and when we can find simplifications of procedure V.2 that keep track of less than players' choices and the full information encoded in belief hierarchies to characterize common belief in rationality. This is clearly possible in traditional games where *iterated elimination of strictly dominated choices* does provide a characterization of common belief in rationality that only explicitly keeps track of players' choices. In section VI we generalize the

result for traditional games by providing a procedure called *iterated elimination of choices and nth-order beliefs* that characterizes common belief in rationality in so-called *belief-finite psychological games* while only keeping track of choices and finite levels of higher-order beliefs.

V.C Example

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.

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 set

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

– 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

$$\bigcap_{k \in \mathbb{N}} R_1(k) \subseteq \varnothing \text{ and } \bigcap_{k \in \mathbb{N}} R_2(k) \subseteq \varnothing.$$

In agreement with our impossibility result from example IV.3, iterated elimination of choices and belief hierarchies yields an empty reduction. Note that $R_i(k) \neq \varnothing$, $i \in \{1,2\}$ for any finite k.[9] 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. 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 an important special case of psychological games. In *belief-finite games*, the utilities of players only depend on finitely many levels of higher-order beliefs:

Definition VI.1. *(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 not hard to see that every belief-finite game preserves rationality at infinity: Suppose that, for belief hierarchy b_i and every $m \geq 1$, there is some $\hat{b}_i$ with $\hat{b}_i^m = b_i^m$ such that choice c_i is rational given $\hat{b}_i$. Now let utility depend on at most nth-order beliefs for some fixed $n \geq 1$. By assumption, there is $\hat{b}_i$ with $\hat{b}_i^n = b_i^n$ such that choice c_i is rational given $\hat{b}_i$. But then, also c_i is rational for b_i. This leads to the following observation:

Observation VI.2. *(Belief Finiteness and Preservation of Rationality at Infinity)*
Every belief-finite game 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.

More interestingly, belief-finite games also allow for a considerably simpler characterization of common belief in rationality. In the remainder of this section, we introduce and study this procedure, which we call *Iterated Elimination of choices and nth-order beliefs*.[10] We will show

[9] 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.

[10] 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.

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. Subsequently, we illustrate *Iterated Elimination of choices and nth-order beliefs* by means of an example. Lastly, we zoom in on an additional issue that can arise if a belief-finite game is not *belief-continuous*: Choices in these games might be rationalizable under *up to k-fold belief in rationality* for every finite k while not being rationalizable under *common belief in rationality*.

VI.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.3. *(Iterated Elimination of Choices and nth-Order Beliefs)*

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

$$R_i^n(1) = \{(c_i, b_i^n) \in C_i \times B_i^n \mid \exists b_i^{n+1} \in B_i^{n+1} \text{ with } \operatorname{marg}_{X_i^n} b_i^{n+1} = b_i^n \text{ such that } u_i(c_i, b_i^{n+1}) \geq u_i(c_i', b_i^{n+1}), \forall c_i' \in C_i\}.$$

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

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

We finally define:

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

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. This naturally generalizes the characterization of common belief in rationality in traditional games: Whenever utility depends on at most $n+1$th-order beliefs, we may eliminate amongst choices and nth-order beliefs. So, in particular, when utility depends only on first-order beliefs, we can resort to the familiar procedure *iterated elimination of strictly dominated choices*.

In a psychological game where utilities depend only on up to $n+1$th-order beliefs, our more general procedure is always 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. As it turns out, though, 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.4. *(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.5:

Theorem VI.5. *(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 $b_i^n = \text{marg}_{X_i^n} b_i^{n+1}$.

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 $(c_i, b_i^n) \in R_i^n(k+2)$, establishing the first direction.

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

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

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 VI.8 further below: 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 VI.8. 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 VI.8, $B_i(k, c_i)$ 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 of choices at the limit of common belief in rationality: 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.

In section VI.D we mount a thorough investigation of *elimination at the limit*, providing a formal proof that *belief continuity* is sufficient to ensure that this phenomenon cannot occur and an explicit example of a *belief-discontinuous belief-finite* game where choices do get eliminated at the limit of common belief in rationality.

We can still find 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. Details and proofs are provided in appendix B.

VI.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. $R_y^1(1) = \{(ditch, b_y^1) | b_y^1(ditch) = 1\} \cup \{(date, b_y^1) | b_y \in B_y^1\}$ and
   $R_a^1(1) = \{(date, b_a^1) | b_a^1(date) \geq \frac{1}{2}\} \cup \{(ditch, b_a^1) | b_a^1(date) \leq \frac{1}{2}\}$.
2. $R_y^1(2) = \{(date, b_y^1) | b_y^1 \in B_y^1\}$ and $R_a^1(2) = R_a^1(1)$.
3. $R_y^1(3) = R_y^1(2)$ and $R_a^1(3) = \{(date, b_a^1) | b_a^1(date) = 1\} =: \{(date, date)\}$.
4. $R_y^1(4) = \{(date, b_y^1) | b_y^1(date) = 1\} =: \{(date, date)\}$ and $R_a^1(4) = R_a^1(3)$.

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.

VI.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,[11] 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 to his choice c_i. Similar to the previous example, we can represent Alice's and your preferences by two pairs of finite matrices contain-

[11] 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).

ing the utilities that you and Alice derive from your extreme first-order beliefs and your extreme second-order expectations. This is shown in table 4 below.

Table 4: The Nightly Encounter

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

+

	ε_y^2	
You	*stay*	*run*
stay	2	0
run	0	3

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

+

	ε_a^2	
Alice	*stay*	*run*
stay	3	0
run	0	2

The total utility for you is then the sum of these two utility components. For instance, your utility from choosing *stay* if your first-order belief 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.

As we will see, *iterated elimination of choices and first-order beliefs* does not terminate within finitely many steps here. The intuition behind the result goes as follows:
You have an inherent preference for choosing *run* over *stay*, so *stay* can only be rationalized for you if you are sufficiently sure that Alice chooses *stay* and/or that she expects you to choose *stay*. In particular, your preference for *run* is so strong that no expectation that Alice might have regarding your choice could make you choose *stay* if you assign full probability to her choosing *run*. So there is a *minimum probability* with which you must think that Alice chooses *stay* in order to rationally choose *stay* yourself. At this minimum probability you are just indifferent between choosing *run* and *stay*, provided you assign full probability to Alice expecting you to choose *stay* in your second-order expectation. By the same reasoning, Alice's preference for *stay* implies that there is a minimum probability that she must assign to you choosing *run* so that she can rationally choose *run* and this minimum probability must then go together with her assigning full probability to you expecting her to choose *run*.

Now assume that you rationally choose *stay* while believing in Alice's rationality. Then, by the preceding reasoning, you must assign some minimum probability to Alice choosing *stay*. Moreover, since you believe Alice to choose rationally, for each probability mass you put on Alice choosing *run*, you have to assume that Alice expects you to choose *run* with the minimum probability that would be necessary to make choosing *run* rational for her. So for each probability mass you put on Alice choosing *run* in your first-order belief, your second-order expectation has to put at least this minimum probability on Alice expecting you to choose *stay*. Consequently, if you believe in

Alice's rationality and you believe her to choose *run* with positive probability, you *cannot* anymore assign full probability to her expecting you to choose *stay* in your second-order expectation and, as a consequence, the minimum probability you have to assign to Alice choosing *stay* so that you can rationally choose *stay* while believing in Alice's rationality will be *strictly higher* than the minimum probability from the preceding step. The same reasoning, mutatis mutandis, implies that Alice must assign a strictly higher minimum probability than before to you choosing *run* so that she can rationally choose *run* and also believe in your rationality.

But then, if you want to choose *stay* under up to 2-fold belief in rationality, you will have to take into account Alice's new minimum probability on you choosing *run* in your second-order expectation and this, in turn, will increase the minimum probability you must put on her choosing *stay* even further.

Continuing in this fashion, it can be shown that, at every level k of up to k-fold belief in rationality, you have to assign a strictly higher minimum probability to Alice choosing *stay* in order to rationally choose *stay* than at the preceding level and similarly for Alice. Consequently, *iterated elimination of choices and first-order beliefs* will take infinitely many steps to converge in this game.

To show this result more formally, we will now explicitly apply *iterated elimination of choices and first-order beliefs* to determine the combinations of choices and first-order beliefs for you and Alice that are consistent with common belief in rationality. Since utility functions here depend linearly on first-order beliefs and second-order expectations, we can conveniently capture elimination steps as linear restrictions on the product space of first-order beliefs and second-order expectations for both you and Alice.[12] To determine 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 $b_y^1(run) \leq \frac{4}{5}$. 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

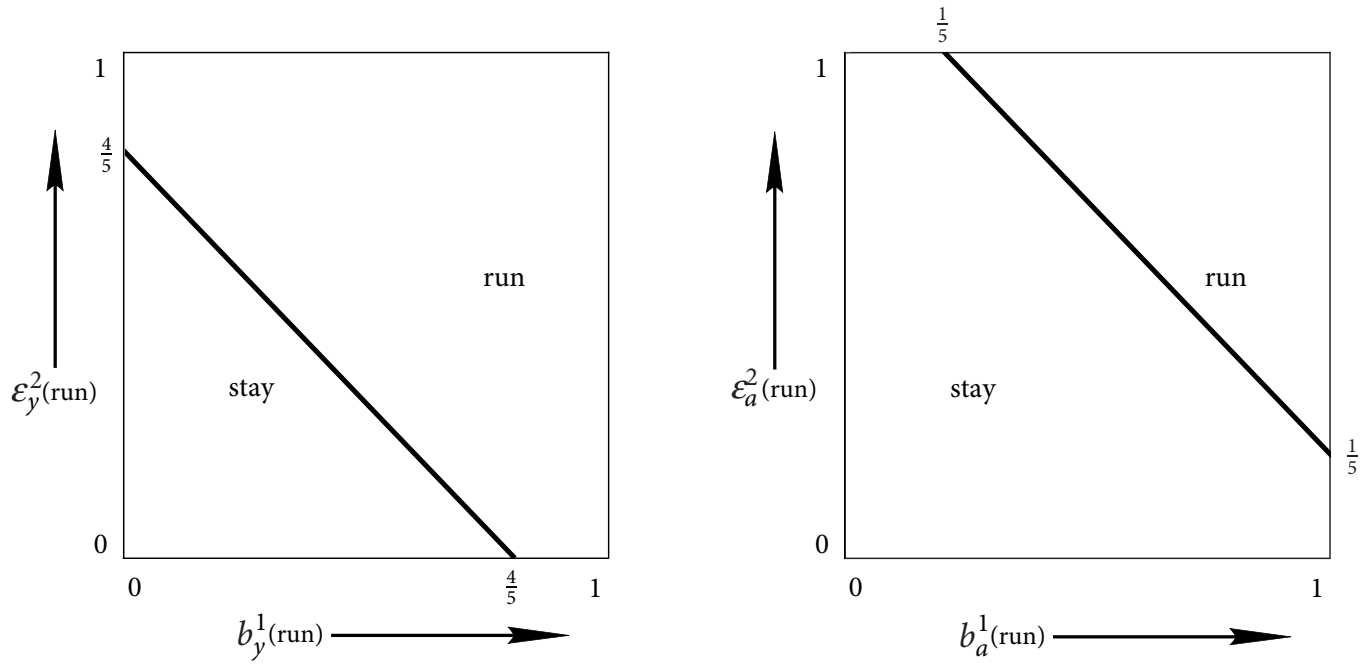
$$R_y^1(1) = \left\{(stay, b_y^1) \middle| b_y^1(run) \leq \frac{4}{5}\right\} \quad \cup \quad \{(run, b_y^1) | b_y^1 \in \Delta(\{stay, run\})\}.$$

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

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

[12] This special property is more generally true of additive expectation-based games and can be used to define a simplified version of iterated elimination of choices and nth-order beliefs. See our companion paper Jagau and Perea (2017).

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 $\text{Conv}(R_a^1(1))$ 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 $\text{Conv}(R_a^1(1))$ for which *run* is rational. Hence, we obtain that

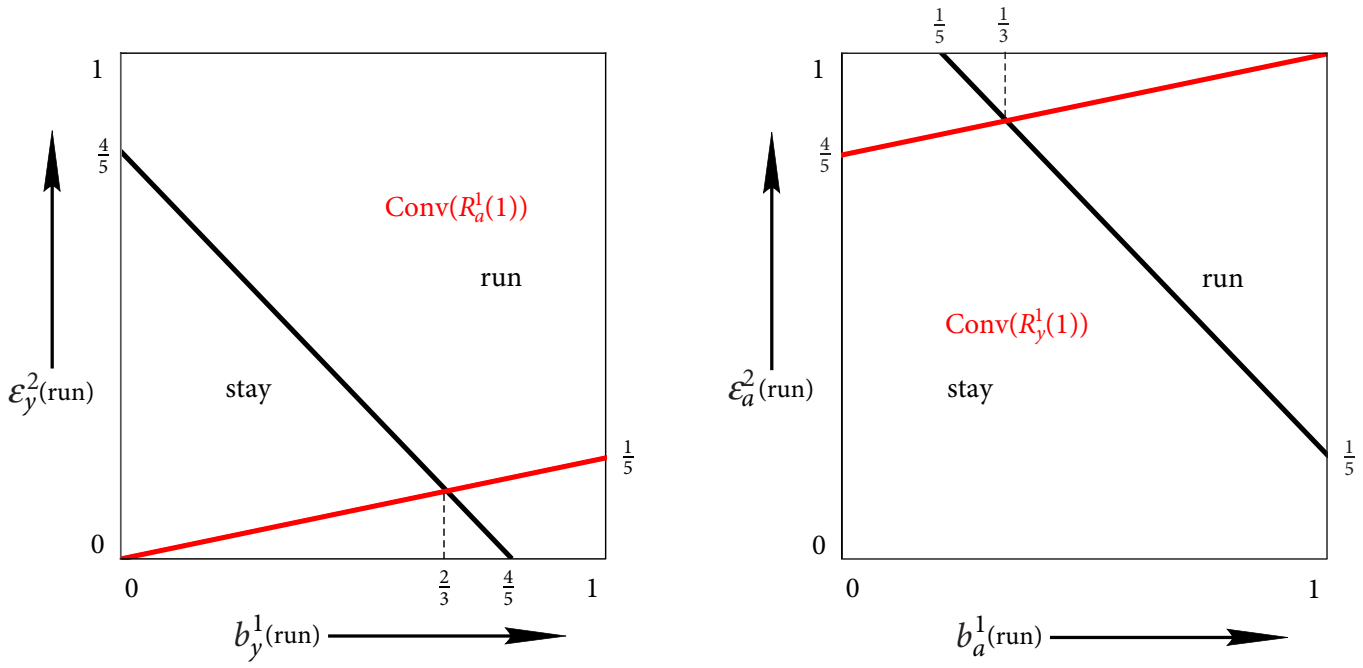
$$R_y^1(2) = \left\{(stay, b_y^1) \middle| b_y^1(run) \leq \frac{2}{3}\right\} \quad \cup \quad \{(run, b_y^1) | b_y^1 \in \Delta(\{stay, run\})\}.$$

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 *iterated elimination of choices and first-order beliefs* does not terminate within finitely many steps.

Figure 2: Convex hull of $R_a^1(1)$ and $R_y^1(1)$



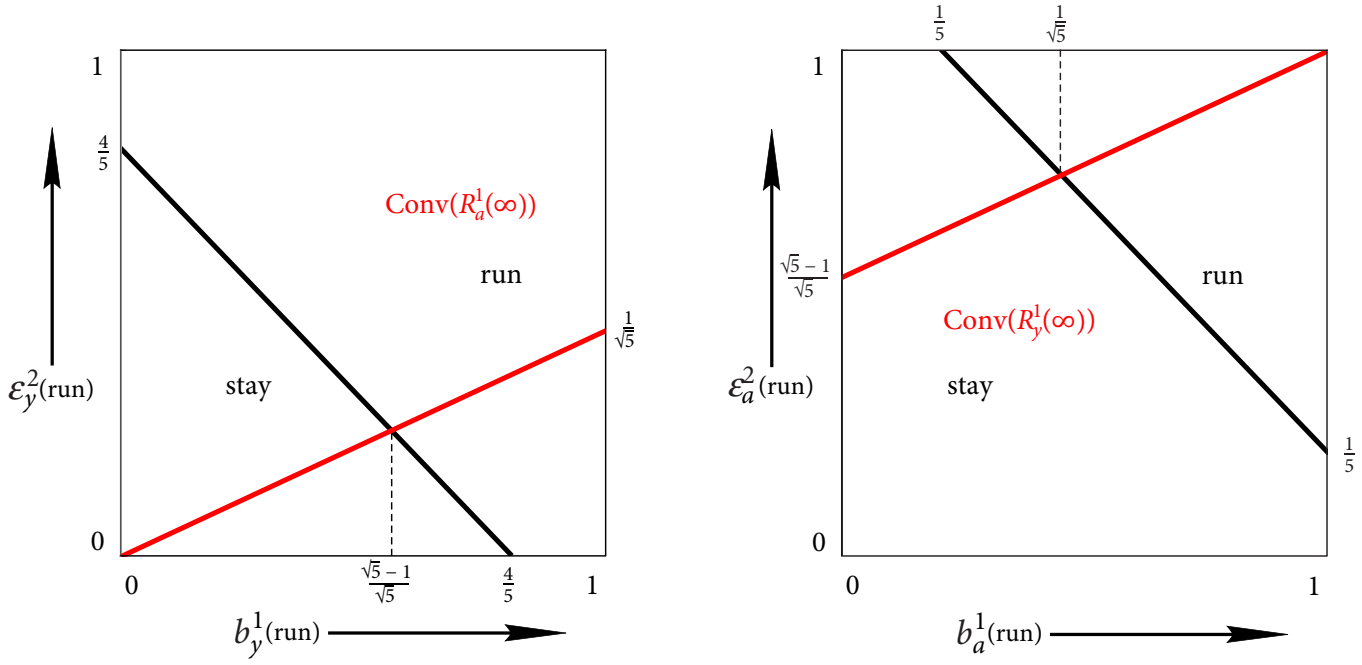
Finally, it can be verified that

$$R_y^1(\infty) = \left\{(stay, b_y^1) \middle| b_y^1(run) \leq \frac{\sqrt{5}-1}{\sqrt{5}}\right\} \quad \cup \quad \{(run, b_y^1) | b_y^1 \in \Delta(\{stay, run\})\} \text{ and}$$

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

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 $R_y^1(\infty)$ and $R_a^1(\infty)$ can be graphically constructed.

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



VI.D Elimination at the Limit

Even though belief-finite psychological games trivially *preserve rationality at infinity* and therefore the possibility of common belief in rationality is always guaranteed within this class of games, they can still exhibit an 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. 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.[13] Interestingly, *preservation of rationality at infinity* is therefore only enough to ensure the possibility of common belief in rationality, but not the closedness of the *belief in the opponent's rationality*-operator. This was precisely the reason why our procedure VI.3 only yielded a tight characterization of common belief in rationality for *belief-continuous* belief-finite games.

In this section we take a closer look at elimination at the limit. First, we show that 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.

[13] 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 the present subsection, we examine this phenomenon more closely.

Next, we provide an example of a belief-discontinuous game that preserves rationality at infinity and where we can indeed *eliminate choices at the limit* of common belief in rationality.

Theorem VI.8. *(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 $B_i(0, c_i), B_i(1, c_i), \ldots$ 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

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

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. □

An interesting follow-up question to theorem VI.8 is whether we can indeed find a *belief-discontinuous game* that *preserves rationality at infinity* but where some choice gets *eliminated at the limit* of common belief in rationality. In the following example, we construct such a game by slightly modifying the game from example VI.7.

Example VI.9. (Elimination of Choices at the Limit of Common Belief in Rationality)
Betting on Alice's Rationality: (inspired by Dufwenberg and Stegeman 2002)
Consider once more the Nightly Encounter from example VI.7 and let Alice's and your preferences be exactly as we defined them there. Now assume that a third player, *Bob*, is watching the scene from his bedroom window. Bob has a long-standing interest in epistemic game theory, so one of the first questions that pops up in his mind is whether Alice entertains common belief in rationality. Specifically, having analyzed the situation up to the point that we analyzed it in example VI.7, Bob wants to guess whether Alice does entertain common belief in rationality or not. To concisely write down Bob's preferences, let $b^2_{b,a} = \text{marg}_{C_a \times B^1_a} b^2_b$. In words, $b^2_{b,a}$ is Bob's second-order belief regarding only Alice. Bob's preferences are then described by the following matrix:

Table 5: Betting on Alice's Rationality

Bob	$b^2_{b,a} \in \Delta(R^1_a(\infty))$	$b^2_{b,a} \notin \Delta(R^1_a(\infty))$
CBR	1	0
No CBR	0	1

Betting on Alice's Rationality is still a belief-finite game and therefore it preserves rationality at infinity. However, Bob's utility function is not belief-continuous. For example, perturbing Bob's second-order beliefs about Alice's choice and first-order belief slightly around the degenerate belief that assigns full probability to $c_a = run, b^1_a(run) = \frac{1}{\sqrt{5}}$ can make $u_b(CBR, b^2_b)$ jump discontinuously from 1 to 0.

It is now easily verified that Bob's choice $No\,CBR$ can be eliminated, but only at the limit of common belief in rationality: To see this, recall that $R^1_a(k) \supsetneq R^1_a(\infty)$ for all finite k as we saw in example VI.7. So at any finite level k of up to k-fold belief in rationality, Bob's second-order belief can assign full probability to combinations of choices and first-order beliefs for Alice that lie outside $R^1_a(\infty)$. At common belief in rationality, however, Bob's second-order belief must assign full probability to the set $R^1_a(\infty)$. It follows that Bob can rationally choose $No\,CBR$ under up to k-fold belief in rationality for any finite k, but not under common belief in rationality.

The preceding example of elimination at the limit is reminiscent of examples using traditional games in Lipman (1994), Dufwenberg and Stegeman (2002), and Bach and Cabessa (2012).[14] Different from our example, however, all traditional-game examples involve infinite choice sets.

As pointed out above, the possibility of elimination at the limit implies that our algorithm *iterated elimination of choices and nth-order beliefs* (procedure VI.3) does not tightly characterize common belief in rationality in *belief-finite* games that are not *belief-continuous*. In appendix B, we show that this issue can be dealt with, but at the cost of using a much more cumbersome

[14] In particular, the structure of Bob's decision problem is borrowed from Dufwenberg and Stegeman's (2002) example 3 where, in a 3-player game, player 3 observes a Cournot competition between players 1 and 2 and bets on whether the unique rationalizable outcome of the game between 1 and 2 will materialize or not.

algorithm which we call *iterated elimination of choices and nth-order and higher-order beliefs.*

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 eliminations 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 have traditional outcome-based preferences that 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.

Many examples of static psychological games that have been studied actually are unilateral psychological games (cf. Geanakoplos et al. 1989, Kolpin 1992).[15]

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.

[15] Also, a lot of dynamic psychological games have a unilateral structure (examples are in Huang and Wu 1994, Dufwenberg 2002, Dufwenberg and Kirchsteiger (2004), Charness and Dufwenberg 2006, Falk and Fischbacher (2006), Battigalli and Dufwenberg 2007, Battigalli and Dufwenberg 2009). It is not hard to see that the finiteness result we present here would readily extend to all such examples after translating our definitions to the dynamic psychological games framework.

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*

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

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

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

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.*

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

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

Given this simplification, we can now easily see that $\text{proj}_{C_1} R_1^1(K) = \text{proj}_{C_1} R_1^1(K+1)$ and $\text{proj}_{C_i} R_i^1(K) = \text{proj}_{C_i} R_i^1(K+1)$, $i \neq 1$ 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 $\text{proj}_{C_i} R_i^1(K+1) = \text{proj}_{C_i} R_i^1(K+2)$, $i \neq 1$ and, by assumption, $\text{proj}_{C_1} R_1^1(K+1) = \text{proj}_{C_1} R_1^1(K+2)$, 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* (see appendix B) which is not in general a finite procedure.

VIII Discussion

VIII.A Alternative Modeling Approaches

A reader familiar with the existing psychological-games literature will likely have noticed that in definition II.1 we model static psychological games slightly differently than previous contributions. The best-known modeling approaches in the previous 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 is mathematically equivalent to the alternative definitions used in these papers. Here, we show the equivalence for the vastly more popular framework from Battigalli and Dufwenberg (2009). In appendix C, we do the same for the one from Geanakoplos et al. (1989).

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 $\times_{j\in I} C_j$, 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

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

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

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

where b_i is identified with a probability measure over $C_{-i} \times B_{-i}$. 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, for static psychological games, this modeling approach can be mapped into our framework and vice versa. Moreover, we see that dependence on opponents' choices and belief hierarchies is actually redundant. As is reasonable, a player i in Battigalli and Dufwenberg's (2009) framework forms beliefs about these objects in light of his belief hierarchy b_i. So appearances notwithstanding, allowing utility to depend on (c_{-i}, b_i, b_{-i}) is just as general as making it depend only on b_i.

Form a conceptual point of view, our approach can be said to differ from previous ones 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.

Battigalli and Dufwenberg (2009) (and also Geanakoplos et al. 1989) assume an *observers' perspective*: 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 seem 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 beliefs expressing coherency and common belief in coherency 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 pro-

ceed 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, whether the *one-person* or the *observers'* approach to modeling games 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). However, as we saw when comparing the modeling framework from Battigalli and Dufwenberg (2009) to ours, the *observers approach* leads to redundancies in the function arguments of players' utilities which might easily cause confusion. For example, the expectation over $C_{-i} \times B_{-i}$ that is taken in expression (1) might suggest that belief-dependent preferences in their framework preserve the linearity in first-order beliefs that we are used to from traditional games. Clearly though, the direct dependence of expression (1) on the belief hierarchy b_i still allows for arbitrary non-linearities in belief-dependent preferences so that the expectational term eventually carries no informational content regarding the shape of preferences.

VIII.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 short of what we would be used to from traditional games.

In this paper we started theorizing at the most fundamental level 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), this was done for pedagogical and not for conceptual reasons. We are confident that our results carry over to this richer class after appropriate modifications of the definitions. Since we restricted our analysis to common belief in rationality which, different from concepts for reasoning in dynamic games, assumes no restrictions on how players update their beliefs while a dynamic game unfolds, it seems clear that allowing for sequential interaction and dependence of utility on updated beliefs cannot lead to qualitative differences for our investigation. The computational issues raised by us here will arise in the same fashion in dynamic psychological games and, if anything, only raise the complexity bar relative to static psychological games.

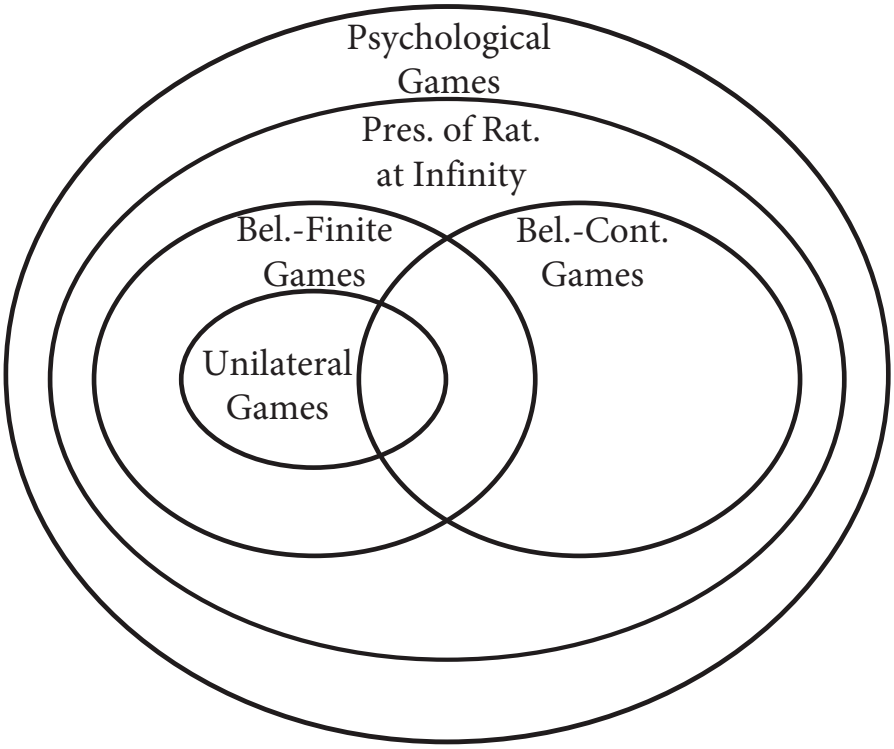
Our results not only relax the previously known existence conditions for common belief in rationality in psychological games, but also provide iterative procedures that select the choices that can be made under common belief in rationality in a given psychological game. As we saw,

special classes of psychological games allow for massive simplifications in these algorithms relative to the more general case. 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	Algorithm for Common Belief in Rationality
All Games (Definition II.1)	Not guaranteed (example IV.3)	Not guaranteed (example IV.9)	Elimination of choices and belief hierarchies (procedure V.2)
Preservation of Rationality at Infinity (Definition IV.1)	Guaranteed (theorem IV.2)	⋮	⋮
Belief-Continuous Games (Definition IV.4)	⋮	Guaranteed (cf. Geanakoplos et al. 1989)	⋮
Belief-Finite Games (Definition VI.1)	⋮	Not guaranteed (example IV.9)	Elimination of choices and nth- and higher-order beliefs (procedure B.1, in appendix)
Belief-Finite, Belief-Continuous Games	⋮	Guaranteed (cf. Geanakoplos et al. 1989)	Elimination of choices and nth-order beliefs (procedure VI.3)
Belief Continuous, Unilateral Games (Definitions IV.4,VII.1)	⋮	⋮	Elimination of choices and 1st-order beliefs (finite by theorem VII.2)

Figure 4: Classes of Psychological Games



Appendix

A Construction of Beliefs

Following Brandenburger and Dekel (1993), 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

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

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 A.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 $\tilde{B}_i(1) \subset \tilde{B}_i(0)$ 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. B_i should be the set of i's beliefs 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

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

and $B_i = \bigcap_{k\geq 0} \tilde{B}_i(k)$.

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 $B_i^n = \text{proj}_{\Delta(X_i^n)} B_i$.

B An Algorithm for General Belief-Finite Games

Here we 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 B.1. *(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:

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

Theorem B.1. *(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 $\text{proj}_{C_i \times B_i^n} R_i^{n\uparrow}(\infty)$.

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.3) 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

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

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. □

C Geanakoplos et al.'s (1989) Approach to Psychological Games

In Geanakoplos et al. (1989) definition of static psychological games, 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.1.

Bibliography

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

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, IGIER Working Paper No. 506, Link: http://www.igier.unibocconi.it/files/506.pdf.

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

Battigalli, P., 1997: On rationalizability in extensive games. *Journal of Economic Theory*, **74 (1)**, 40–61.

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, 2017: Frustration and anger in games, IGIER Working Paper No. 539, Link: http://www.igier.unibocconi.it/files/539.pdf.

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, 2013: Language-based games. *Proceedings of Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge*.

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.

Dekel, E., D. Fudenberg, and D. K. Levine, 1999: Payoff information and self-confirming equilibrium. *Journal of Economic Theory*, **89 (2)**, 165–185.

Dekel, E., D. Fudenberg, and D. K. Levine, 2002: Subjective uncertainty over behavior strategies: A correction. *Journal of Economic Theory*, **104 (2)**, 473–478.

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. Stegeman, 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, University of Arizona Working Paper 16-05, Link: https://econ.arizona.edu/sites/econ/files/wp2016-05-liesindisguise_november6-2016.pdf.

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

Fudenberg, D., and D. K. Levine, 1983: Subgame-perfect equilibria of finite and infinite horizon games. *Journal of Economic Theory*, **31 (2)**, 251–268.

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 and psychological expected utility, 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.

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.

Perea, A., 2014: Belief in the opponents future rationality. *Games and Economic Behavior*, **83**, 231–254.

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.

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

Tversky, A., and D. Kahneman, 1986: Rational choice and the framing of decisions. *The Journal of Business*, **59 (4)**, 251–278.