Psychological Expected Utility

Stephan Jagau[ABC] and Andrés Perea[BD]

November 1, 2018

Psychological game theory is a rich framework for modeling belief-dependent motivations and emotional mechanisms such as surprise, anger, guilt, and intention-based reciprocity. At the same time, putting psychological game theory to work in behavioral and experimental economics tends to be complicated. Many properties and techniques that make traditional games "nice" do not in general extend to psychological games. Hence it becomes an important issue to figure out which psychological games are tractable enough to be useful for applied work.

In this paper, we identify a large class of psychological games that are almost as tractable as traditional games. In these *expectation-based games*, utility linearly depends on recursively constructed summary statistics of players' higher-order beliefs. We argue that this restriction is the natural generalization of expected utility to psychological games, and we show that it has an attractive epistemological interpretation in terms of a one-theory-per-choice condition. It turns out that our assumptions are compatible with all known applications of psychological games, and that they enable massive simplifications in their analysis. In particular, all commonly studied examples of psychological games are found to be numerically solvable using standard techniques based on linear programming.

JEL classification: C72, D03, D83

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

[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
[D]Email: a.perea@maastrichtuniversity.nl. Web: http://www.epicenter.name/Perea/

I Introduction

From traditional game theory, we are used to deal with decision-makers that 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, not only outcomes, but also beliefs and intentions of ourselves and others do matter for how we choose to act, transcending the outcome-based preferences from traditional game theory. The traditional approach to game theory thus has a hard time trying to capture fundamental aspects of human behavior as diverse as intention-based reciprocity (Rabin 1993, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006, Sebald 2010), guilt (Huang and Wu 1994, Dufwenberg 2002, Charness and Dufwenberg 2006, Battigalli and Dufwenberg 2007, Attanasi et al. 2016, Attanasi et al. 2017), social pressure and conformity (Huck and Kübler 2000, Li 2008), anxiety (Caplin and Leahy 2004), lying aversion (Battigalli et al. 2013, Dufwenberg and Dufwenberg 2016), surprise (Khalmetski et al. 2015), anger (Battigalli et al. 2015), and esteem (Akerlof 2017). *Psychological game theory* (Geanakoplos et al. 1989, Battigalli and Dufwenberg 2009) addresses this issue by allowing players' utilities to directly depend not only on their choices and beliefs about others' choices, but also on arbitrary levels of higher-order beliefs.

Since its introduction, the psychological games framework has proven to be a useful tool for many applications in behavioral and experimental economics, including – but not limited to – the formal and experimental treatments of psychological phenomena mentioned above.

While psychological games are popular, they tend to be very hard to analyze compared to traditional games. In particular, Jagau and Perea (2018) show that many nice properties we normally take for granted, like possibility of common belief in rationality, existence of equilibrium, and finiteness of the algorithms characterizing common belief in rationality do not in general carry over to psychological games and sometimes require very strong assumptions to be recovered (see also to be done). Hence, it becomes a separate issue of interest for applications to determine which psychological games are easy to work with in applications while also remaining rich enough to provide a realistic model of belief-dependent motivation.

In the present paper, we contribute towards this goal by providing the first formal definition of **expectation-based psychological games** in which preferences only depend on certain *summary statistics of higher-order beliefs*. Utility functions of players in these games are shown to naturally generalize the standard expected utility representation, which is why we call them *psychological expected utility functions*.

While a formal definition of expectation-based games and psychological expected utility has been lacking up to now (but see Caplin and Leahy 2001), they already are widely used in the literature in static and dynamic variants (cf. Rabin 1993, Battigalli and Dufwenberg 2007, Dufwenberg and Dufwenberg 2016). In fact, we are not aware of any application of psychological games that does not study a (possibly dynamic) expectation-based game.

Also, we argue that psychological expected utility has an attractive epistemic interpretation

as the requirement that uncertainty across different levels of belief in the reasoning of a given should be commensurable according to a common probabilistic currency. For example, if a player i cares about what his opponents expects him to do (second-order expectation), then it should not matter whether i is uncertain about his opponent's belief regarding i's choice or whether i thinks his opponent belief expresses uncertainty about i's choice. As long as the implied probabilities with which i think his opponent expects him to choose either of his actions are the same, such considerations should be irrelevant for i's preferences over actions.

After introducing expectation-based games and psychological expected utility, we move on to study the most basic mode of reasoning in games, common belief in rationality (Brandenburger and Dekel 1987, Tan and Werlang 1988, characterizing correlated rationalizability), providing an iterative procedure that characterizes common belief in rationality in general expectation-based psychological games (see section V). Similar to what is true of *iterated elimination of strictly dominated choices in traditional games*, we show that common belief in rationality can be characterized by a linear procedure in expectation-based games games where utility depends on finitely many levels of higher-order beliefs. However, different from traditional games, we also show that the procedure is typically not finite even in a particularly well-behaved subclass of expectation-based games.

The results we present here are closely related to our companion paper Jagau and Perea (2018) where we provide a general analysis of common belief in rationality in static psychological games. In the following we will limit ourselves to static psychological games with two players for reasons of clarity. The full paper will also include the (straightforward) extension of our results to general extensive-form psychological games[1] as well as a sketch of modeling choices for n-player psychological games.

The remainder of this paper is structured as follows. Section II introduces the general definitions of psychological games and of common belief in rationality. Section III introduces expectation-based games and the psychological expected utility representation. Section IV concerns the possibility of common belief in rationality in expectation-based games. Section V builds on this by developing an algorithm for common belief in rationality in expectation based games, called *Iterated Elimination of Choices and Expectations.* Section VI concludes.

[1]Note that we do not study refinements for dynamic psychological games but only common belief in rationality. Extending the investigation of dynamic expectation-based games to common belief in future rationality and common strong belief in rationality in dynamic psychological games is an interesting avenue for future research. For general dynamic psychological games, common strong belief in rationality has already been investigated in Battigalli and Dufwenberg (2009).

II Preliminaries

In this section, we introduce the general framework of psychological games and define common belief in rationality within that framework. For a more detailed exposition of the same material we refer to our companion paper Jagau and Perea (2018).

A. Psychological Games

We start by providing 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 bounded 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.[2] In Jagau and Perea (2018) (Appendix A), we illustrate the adaption of their construction to our setup in full detail.

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.

Some infinite belief hierarchies in B_i can conveniently be captured using a finite epistemic model M.

[2] Following their approach, for any polish space S, let $\Delta(S)$ denote the set of probability measures on the Borel-field over S and endow $\Delta(S)$ with the weak topology. In our case, the primitive space of uncertainty for player i is the set of opponents' choices $\times_{j \neq i} C_j = C_{-i}$.

Definition II.2. *(Epistemic Model)*
For every psychological game Γ*, an* **epistemic model** $M(\Gamma) = (T_i, \beta_i)_{i \in I}$ *specifies a set of types for player* i *and a function* $\beta_i : T_i \to \Delta(C_{-i} \times T_{-i})$ *for every player* i*. So every type* $t_i \in T_i$ *is associated with a belief* $\beta_i(t_i)$ *– a probability distribution over the opponents' choice-type combinations. If* T_i *is finite for every player, we call* M *a* **finite epistemic model**.

Starting from an epistemic model, we can easily recover higher-order beliefs from types in the usual way (cf. Heifetz and Samet 1998).

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.

B. 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 II.3. *(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 II.3, 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 II.4. *(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 II.5. *(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 II.6. *(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 $b_i \in B_i(\infty)$.*

III Expectation-Based Games

A. The Expectation-Based-Games Framework

In this section, we build on the general definition of psychological games to formally define the *expectation-based games* that we focus on in our analysis. Before moving on to the general definition of expectation-based games, we consider the following introductory example:

Example III.1. (A Simple Expectation-Based Game)
Positively Surprising Alice:
You have now been seeing Alice for a month and you plan to get her a present. At various points Alice has casually let you know that she would like to get a box of *chocolate* or a bottle of *wine*. You do not have a strong preference for either of the two potential presents. Instead, you would like it most if you surprise Alice, that is if you get the present that she least expects to receive. In addition, Alice can choose to *accept* or *reject* the present and of course you care for surprising Alice only in case she also accepts the present. If she rejects the present, you cannot enjoy the surprise since Alice does not even open the present. So consequently your utility remains at zero.

We model this situation as a two-player game with player set $I = \{y, a\}$, choice set $C_y = \{chocolate,\ wine\}$ for you and choice set $C_a = \{accept,\ reject\}$ for Alice. As noted before, you

would like to choose the present which you think Alice will find most unlikely *in expectation*, provided she will accept it. Formally, for all choices and belief hierarchies that you deem possible for Alice, you want to choose that present among the ones she accepts that is assigned the lowest *expected probability* over all these belief hierarchies. We will refer to this expected probability as your *second-order expectation*. Formally, for a given belief hierarchy b_y, choice c_a for Alice and choice c_y for you, the induced second-order expectation $e_y^2[b_y](c_a, c_y)$ is given by

$$e_y^2[b_y](c_a, c_y) = \int_{\{c_a\} \times B_a} b_a^1(c_y) \, \mathrm{d}b_y.$$

That is, for each choice c_a for Alice and each choice c_y of you, $e_y^2[b_y](c_a, c_y)$ records the *expected probability* with which you think Alice chooses c_a and believes you to choose c_y.

Let your utility function now be

$$u_y(c_y, b_y) = \sum_{c_y' \neq c_y} e_y^2[b_y](accept, c_y').$$

For brevity, we refrain from modeling Alice's preferences over the choices *accept* and $reject$.

To give an example of how these expectations are computed, we construct an epistemic model for Positively Surprising Alice. The epistemic model is given in table 1 below.

Table 1: An Epistemic Model for Positively Surprising Alice

Types	$T_y = \{t_y^1, t_y^2\}$
	$T_a = \{t_a^1, t_a^2\}$
Beliefs for You	$b_y(t_y^1) = (\frac{1}{2} \cdot (accept, t_a^1) + \frac{1}{2} \cdot (reject, t_a^2))$
	$b_y(t_y^2) = (reject, t_a^2)$
Beliefs for Alice	$b_a(t_a^1) = (\frac{2}{3} \cdot (wine, t_y^2) + \frac{1}{3} \cdot (chocolate, t_y^2))$
	$b_a(t_a^2) = (chocolate, t_y^2)$

We can now easily compute the second-order expectation for each type of you in the model. For example, for Alice's choice *accept*, your choice *chocolate*, and your type t_y^1, we compute:

$$\begin{aligned} e_y^2[b_y(t_y^1)](accept, chocolate) &= \int_{\{accept\} \times B_a} b_a^1(chocolate) \, \mathrm{d}b_y(t_y^1) \\ &= \frac{1}{2} b_a^1(t_a^1)(chocolate) \\ &= \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \end{aligned}$$

Hence $u_y(wine, b_y(t_y^1)) = \frac{1}{6}$ and similarly it may be verified that $u_y(chocolate, b_y(t_y^1)) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$.

The "expected probability" or *second-order expectation* we use here may be viewed as a sum-

mary statistic of your second-order belief, which in this case is sufficient to determine your utility. Moreover, your utility in Positively Surprising Alice depends *linearly* on this summary statistic.

We now give the formal definition of expectation-based games. First, we need to make it clear how *expectations* can be derived from a *belief hierarchy*. To keep things tractable, we restrict to two-player games. Consider a belief hierarchy b_i for player i. The *first-order expectation* $e_i^1[b_i]$ induced by b_i is simply the first-order belief induced by b_i. That is,

$$e_i^1[b_i] := b_i^1,$$

which is a probability distribution on C_j. Starting from first-order expectations, we recursively define k-th order expectations, for every $k \geq 2$.

Before giving the general definition, we develop some intuitions using second-order expectations: Like the first-order expectation e_i^1, i's second-order expectation e_i^2 should collect the probabilities i assigns to the opponent's choices c_j. In addition, to each of these choices c_j and each of i's choices c_i it should match the average probability i assigns to his opponent believing that he chooses c_i given that his opponent chooses c_j. So e_i^2 becomes a *joint probability measure* on $C_j \times C_i$:

$$e_i^2[b_i](c_j, c_i) := \int_{\{c_j\}\times B_j} e_j^1[b_j](c_i)\, \mathrm{d}b_i = \int_{\{c_j\}\times B_j} \int_{\{c_i\}\times B_i} \mathrm{d}b_j \mathrm{d}b_i$$

Iterating this construction, we arrive at the general definition of expectations. For ease of notation, we recursively define

$$C_i^1 := C_j \text{ and, for all } k > 1,\ C_i^k := \begin{cases} C_i^{k-1} \times C_i, & k \text{ is even} \\ C_i^{k-1} \times C_j, & k \text{ is odd} \end{cases} = \begin{cases} \overbrace{C_j \times C_i \times \ldots C_i}^{k \text{ times}}, & k \text{ is even} \\ \underbrace{C_j \times C_i \times \ldots C_j}_{k \text{ times}}, & k \text{ is odd} \end{cases}.$$

where, for each $k \geq 1$, C_i^k will become the domain of player i's kth-order expectation. Representatives of the set C_i^k will be denoted by c_i^k. Also, with some abuse of notation, we will often identify $c_i^k = (c_j, c_j^{k-1})$.

We are now ready to define higher-order expectations:

Definition III.2. *(kth-Order Expectation)*
Let $\Gamma = (C_i, B_i, u_i)_{i\in I}$ be a two-player psychological game with $I = \{i, j\}$ and let b_i be a belief hierarchy for player i. Define $e_i^1[b_i] := b_i^1$. For $k \geq 2$, the **kth-order expectation** *of player i given b_i is now defined to be*

$$e_i^k[b_i](c_i^k) = e_i^k[b_i](c_j, c_j^{k-1}) := \int_{\{c_j\}\times B_j} e_j^{k-1}[b_j](c_j^{k-1})\, \mathrm{d}b_i, \text{ for every } c_i^k \in C_i^k$$

where in the integral we use b_i as a probability measure on $C_j \times B_j$.

For every $k \geq 1$, let

$$E_i^k := \{e_i^k[b_i] \mid b_i \in B_i\} = \Delta(C_i^k)$$

be the set of all k-th order expectations for player i. Further, by

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

we will denote the *vector of expectations* induced by the belief hierarchy b_i and

$$E_i = \{e_i[b_i] \mid b_i \in B_i\} = \left\{(e_i^1, e_i^2, \dots) \in \bigtimes_{k\geq 1} E_i^k \mid \text{marg}_{C_i^k} e_i^{k+1} = e_i^k,\ k \geq 1\right\}$$

will denote the set of all expectation vectors.

It is intuitively clear that the induced vector of expectations $e_i[b_i]$ depends *linearly* on the belief hierarchy b_i. To formalize this statement, consider two belief hierarchies b_i and b_i' for player i, and a number λ between 0 and 1. Then, by the convex combination $\lambda b_i + (1-\lambda)b_i'$ we denote the belief hierarchy $\hat{b}_i$ given by

$$\hat{b}_i(\{c_j\} \times \hat{B}_j) := \lambda b_i(\{c_j\} \times \hat{B}_j) + (1-\lambda) b_i'(\{c_j\} \times \hat{B}_j)$$

for all $c_j \in C_j$ and all measurable subsets $\hat{B}_j \subseteq B_j$. Then, it may be checked that

$$e_i^k[\lambda b_i + (1-\lambda)b_i'] = \lambda e_i^k[b_i] + (1-\lambda) e_i^k[b_i']$$

for all $k \geq 1$, all belief hierarchies b_i and b_i', and all $\lambda \in [0,1]$. Hence, the induced vector of expectations depends linearly on the belief hierarchy.

Similar to how a belief hierarchy $(b_i^1, b_i^2, \dots)$ can be identified with a unique probability measure on the product space $C_j \times B_j$ of opponent's choices and belief hierarchies, we can show that a vector of expectations $e_i[b_i]$ corresponds to a unique probability measure on $C_i^\infty = \times_{k\geq 1}(C_j \times C_i)$ where that measure also preserves the linearity of the induced expectations $e_i^k[b_i]$.

Theorem III.3.
There is a linear homeomorphism $f : E_i \to \Delta(C_i^\infty)$ such that $\text{marg}_{C_i^k} f(e_i) = e_i^k$ for all $k \geq 1$.

Proof.
Take some $e_i[b_i] \in E_i$ where b_i is some inducing belief hierarchy. To prove the result, we first show that $e_i^k[b_i] = \text{marg}_{C_i^k} e_i^{k+1}[b_i]$, for all $k \geq 1$. As follows from Lemma 1 in Brandenburger and Dekel (1993), there then exists a homeomorphism $f : E_i \to \Delta(C_i^\infty)$ such that $f(e_i[b_i])$ is the unique distribution over C_i^∞ with $\text{marg}_{C_i^k} f(e_i[b_i]) = e_i^k[b_i]$ for all $k \geq 1$.

Induction Start: For $k = 1$, we have $C_i^1 = C_j$, $C_i^2 = C_j \times C_i$, and

$$\begin{aligned}
\operatorname{marg}_{C_j} e_i^2[b_i](c_j) &= \sum_{c_i \in C_i} e_i^2[b_i](c_j, c_i) \\
&= \sum_{c_i \in C_i} \int_{\{c_j\} \times B_j} e_j^1[b_j](c_i)\, \mathrm{d}b_i \\
&= \int_{\{c_j\} \times B_j} \sum_{c_i \in C_i} e_j^1[b_j](c_i)\, \mathrm{d}b_i \\
&= \int_{\{c_j\} \times B_j} \mathrm{d}b_i \\
&= e_i^1[b_i](c_j)
\end{aligned}$$

Induction Step: For $k+1$, take some $c_i^{k+1} = (c_j, c_j^k)$ and assume that k is even.[3] Assume that, for all $k \geq 1$ and all players i, $e_i^k[b_i](c_i^k) = \operatorname{marg}_{C_i^k} e_i^{k+1}[b_i](c_i^k)$. We then have

$$\begin{aligned}
\operatorname{marg}_{C_i^{k+1}} e_i^{k+2}[b_i](c_i^{k+1}) &= \sum_{c_i \in C_i} e_i^{k+2}[b_i](c_i^{k+1}, c_i) \\
&= \sum_{c_i \in C_i} \int_{\{c_j\} \times B_j} e_j^{k+1}[b_j](c_j^k, c_i)\, \mathrm{d}b_i \\
&= \int_{\{c_j\} \times B_j} \sum_{c_i \in C_i} e_j^{k+1}[b_j](c_j^k, c_i)\, \mathrm{d}b_i \\
&= \int_{\{c_j\} \times B_j} \operatorname{marg}_{C_j^k} e_j^{k+1}[b_j](c_j^k)\, \mathrm{d}b_i \\
&= \int_{\{c_j\} \times B_j} e_j^k[b_j](c_j^k)\, \mathrm{d}b_i \\
&= e_i^{k+1}[b_i](c_j, c_j^k) \\
&= e_i^{k+1}[b_i](c_i^{k+1}),
\end{aligned}$$

where in the second to last step we used the induction assumption.

This completes the induction.

To see that f is linear, take some $\lambda \in [0,1]$ and set $\hat{e}_i = \lambda e_i + (1-\lambda) e_i' \in E_i$. Note that $\hat{e}_i^k = \lambda e_i^k + (1-\lambda)(e')_i^k = \lambda \operatorname{marg}_{C_i^k} f(e_i) + (1-\lambda) \operatorname{marg}_{C_i^k} f(e_i') = \operatorname{marg}_{C_i^k} (\lambda\, f(e_i) + (1-\lambda) f(e_i'))$ for all $k \geq 1$. Since $f(\hat{e}_i)$ is the unique measure such that $\operatorname{marg}_{C_i^k} f(\hat{e}_i) = \hat{e}_i^k$, for all $k \geq 1$, we therefore must also have $f(\hat{e}_i) = \lambda\, f(e_i) + (1-\lambda) f(e_i')$. □

Using f, we can identify every $e_i \in E_i$ with a unique probability distribution $f(e_i) \in \Delta(C_i^\infty)$. Therefore, with a slight abuse of notation we will henceforth use e_i to refer both to the sequence $(e_i^1, e_i^2, \dots)$ and to the probability distribution that this sequence induces following theorem III.3. Also, similar to our notation for finite levels of expectations, we will use c_i^∞ to refer to representatives of the set C_i^∞.

[3] For odd k we proceed analogously, modulo replacing (c_j^k, c_i) by (c_j^k, c_j) below.

In an expectation-based game, the utility of a player only depends on the vector of expectations induced by his belief hierarchy and it depends *linearly* on the vector of expectations.

Definition III.4. *(Expectation-Based Game)*
A two-player psychological game $\Gamma = (C_i, B_i, u_i)_{i \in I}$ *is* **expectation-based** *if, for both players* i*,*

1. $u_i(c_i, b_i) = u_i(c_i, b_i')$ *whenever* $e_i[b_i] = e_i[b_i']$*,*
2. $u_i(c_i, \lambda b_i + (1-\lambda)b_i') = \lambda u_i(c_i, b_i) + (1-\lambda)u_i(c_i, b_i')$*, for all* $\lambda \in [0,1]$*.*

The first part of definition III.4 allows us to write utility as a function

$$u_i : C_i \times E_i \to \mathbb{R}$$

Together with the second condition, which we may call **belief linearity**, this has several nice consequences. In particular, we can generalize expected utility to expectation-based games:

Corollary III.5. *(Psychological Expected Utility)*
Let $\Gamma = (C_i, B_i, u_i)_{i \in I}$ *be a two-player expectation-based psychological game. Then, for both players* i*, there is a function* $v_i : C_i \times C_i^\infty \to \mathbb{R}$ *such that*

$$u_i(c_i, b_i) = \int_{C_i^\infty} v_i(c_i, c_i^\infty) \, \mathrm{d}e_i[b_i](c_i^\infty)$$

for every $c_i \in C_i$ *and* $b_i \in b_i$*.*

Proof.
For every $c_i^\infty \in C_i^\infty$, take b_i with $e_i[b_i](c_i^\infty) = 1$ and define $v_i(c_i, c_i^\infty) := u_i(c_i, b_i)$. By linearity of $e_i \in \Delta(C_i^\infty)$ (theorem III.3) and by belief-linearity, we can write

$$u_i(c_i, b_i) = \int_{C_i^\infty} v_i(c_i, c_i^\infty) \, \mathrm{d}e_i[b_i](c_i^\infty)$$

as desired. □

An important special case arises when utility depends on finitely many levels of beliefs as in example III.1 above. Formally:

Definition III.6. *(Belief-Finite Game)*
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* $u_i(c_i, b_i) = u_i(c_i, \hat{b}_i)$*.*

If utility in a game only depends on n-th order beliefs, it is easily seen that psychological expected utility can be defined as a weighted sum of extreme n-th order expectations:

Observation III.7. *(Belief-Finite Psychological Expected Utility)*
Take a two-player expectation-based game Γ*. If utility depends on at most* n*th-order beliefs, we can write, for all players* i*,*

$$u_i(c_i, b_i) = \sum_{c_i^n \in C_i^n} e_i^n[b_i](c_i^n) v_i(c_i, c_i^n)$$

with $v_i : C_i \times C_i^n \to \mathbb{R}$*.*

Equivalently, for each player i, there is a $|C_i| \times |C_i^n|$ matrix that represents his utility. Each row of the matrix corresponds to a choice of player i and each column corresponds to one of his extreme nth-order expectations. For instance, in example III.1, your utility is represented by:

Table 2: Positively Surprising Alice

	c_y^2			
You	$(accept, chocolate)$	$(accept, wine)$	$(reject, chocolate)$	$(reject, wine)$
$chocolate$	0	1	0	0
$wine$	1	0	0	0

B. Representative Belief Hierarchies

Clearly, every expectation $e_i[b_i]$ is uniquely determined given the belief hierarchy b_i. The reverse is not true, as illustrated in the following example.

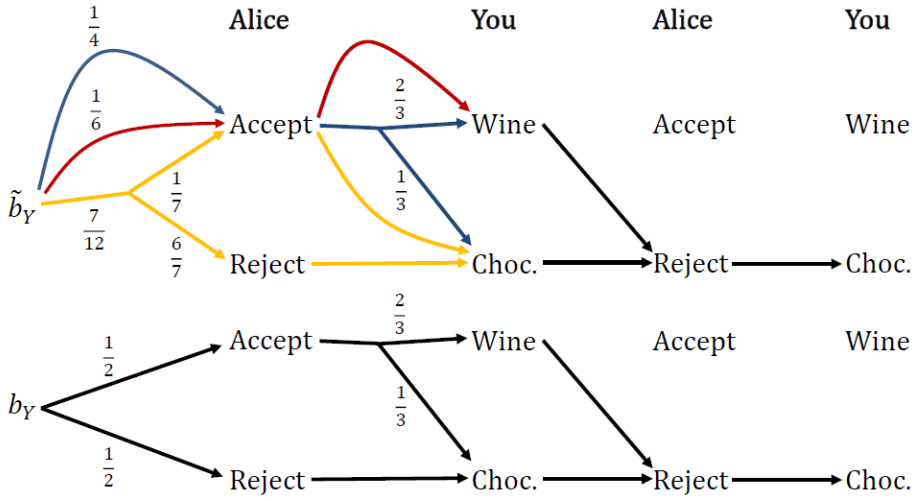
Example III.8. (Expectational Equivalence in the Surprise Game)
Consider the game Positively Surprising Alice from example III.1 above and the extended epistemic model shown below:

Table 3: An Extended Epistemic Model for Positively Surprising Alice

Types	$T_y = \{t_y^1, t_y^2, \tilde{t}_y^1\}$
	$T_a = \{t_a^1, t_a^2, \tilde{t}_a^1\}$
Beliefs for You	$b_y(t_y^1) = (\frac{1}{2} \cdot (accept, t_a^1) + \frac{1}{2} \cdot (reject, t_a^2))$
	$b_y(t_y^2) = (reject, t_a^2)$
	$b_y(\tilde{t}_y^1) = (\frac{1}{4} \cdot (accept, t_a^1) + \frac{1}{6} \cdot (accept, \tilde{t}_a^1) + \frac{7}{12} \cdot ((\frac{1}{7} \cdot accept + \frac{6}{7} \cdot reject), t_a^2))$
Beliefs for Alice	$b_a(t_a^1) = (\frac{2}{3} \cdot (wine, t_y^2) + \frac{1}{3} \cdot (chocolate, t_y^2))$
	$b_a(t_a^2) = (chocolate, t_y^2)$
	$b_a(\tilde{t}_a^1) = (wine, t_y^2)$

Figure 1 below graphically depicts the belief hierarchies b_y and $\tilde{b}_y$ that are induced by your types t_y^1 and $\tilde{t}_y^1$, respectively.

Figure 1: Expectationally Equivalent Belief Hierarchies in the Surprise Game



Looking at the second-order expectation regarding Alice's choice *accept* and your choice *chocolate*, but now for your type $\tilde{t}_y^1$, we can compute

$$\begin{aligned} e_y^2[b_y(\tilde{t}_y^1)](accept, chocolate) &= \int_{\{accept\}\times B_a} b_a^1(chocolate)\, \mathrm{d}b_y(\tilde{t}_y^1) \\ &= \frac{1}{4} b_a^1(t_a^1)(chocolate) + \frac{7}{12} \cdot \frac{1}{7} b_a^1(t_a^2)(chocolate) \\ &= \frac{1}{4} \cdot \frac{1}{3} + \frac{7}{12} \cdot \frac{1}{7} = \frac{1}{6} \end{aligned}$$

So that $e_y^2[b_y(\tilde{t}_y^1)](accept, chocolate) = e_y^2[b_y(t_y^1)](accept, chocolate)$. In a similar fashion, the reader may verify that the induced second-order expectations for t_y^1 and $\tilde{t}_y^1$ coincide for other combinations of choices $(c_a, c_y) \in C_a \times C_y$ and, indeed also for all higher-order expectations. We therefore conclude $e_y[b_y(t_y^1)] = e_y[b_y(\tilde{t}_y^1)]$. That is, t_y^1 and $\tilde{t}_y^1$ are *expectationally equivalent*.

The above example gives us an intuitive idea of what we abstract from in an expectation-based game. As we see, it does not matter whether uncertainty in your belief hierarchy derives from you thinking that Alice is uncertain about your choice in her first-order belief or from you being uncertain about Alice's first-order belief. As long as the induced probabilities of second- and higher-ordered expectations coincide, your utility will be invariant to what source the uncertainty in your model of yours and Alice's strategic reasoning emanates from. Taking expectations, so to say, imposes a common probabilistic currency on the uncertainty encoded in your belief hierarchy at all of its levels.

Given the surjective but not injective relation between the beliefs b_i and the induced expectations $e_i[b_i]$, it makes sense to think of the sets $B_i(e_i) = \{b_i \in B_i | e_i[b_i] = e_i\}$ as defining equivalence

classes on B_i. Each pair of $b_i, b_i' \in B_i(e_i)$ are *expectationally equivalent* and, in an expectation-based game, i's utility function u_i will be invariant on $B_i(e_i)$.

Even though there are typically many belief hierarchies $b_i \in B_i(e_i)$ that induce the same expectation vector e_i, not all of them are created equal. In particular, for each class $B_i(e_i)$ we can identify a "natural" or *canonical* representative $b_i^*[e_i]$ as follows.

Definition III.9. *(Canonical Belief Hierarchy)*
Let $e_i \in \Delta(C_i^\infty)$ be an expectation vector for player i. Now construct the belief hierarchy $b_i^*[e_i]$ as follows:

1. *For all players i, all $e_i \in \Delta(C_i^\infty)$ and all c_j such that $e_i^1(c_j) > 0$, define $e_j^*[e_i, c_j] \in \Delta(C_j^\infty)$ via*
$$e_j^*[e_i, c_j](A) = \frac{e_i(\{c_j\} \times A)}{e_i^1(c_j)}$$
for all measurable $A \subseteq C_j^\infty$.

2. *For both players i, define the set of types $T_i^* = \{t_i^*[e_i] \mid e_i \in \Delta(C_i^\infty)\}$, where*
$$b_i(t_i^*[e_i])(c_j, t_j^*[e_j]) = \begin{cases} e_i^1(c_j), & \text{if } e_i^1(c_j) > 0 \text{ and } e_j = e_j^*[e_i, c_j] \\ 0, & \text{else} \end{cases}$$

3. *For all $e_i \in \Delta(C_i^\infty)$, let $b_i^*[e_i]$ be the belief hierarchy induced by $t_i^*[e_i]$. This is the* **canonical belief hierarchy** *induced by e_i.*

Definition III.9 associates every $e_i \in E_i$ with a unique belief hierarchy $b_i^*[e_i] \in B_i$. To see this, note that condition (1) in definition III.9 selects a unique expectation vector for each e_i and each $c_j \in \text{Supp}(e_i^1)$. Following the recursion defined by (2), $b_i^*[e_i]$ must then identify a unique belief hierarchy in B_i.

Before continuing, we convince ourselves that $b_i^*[e_i]$ indeed induces e_i.

Observation III.10. *$e_i[b_i^*[\hat{e}_i]] = \hat{e}_i$ for all $\hat{e}_i \in E_i$.*

Proof. To prove the statement, show that $e_i^k[b_i^*[\hat{e}_i]] = \hat{e}_i^k$ for all $k \geq 1$. We proceed by induction over $k \geq 1$.

Induction Start: For $k = 1$, we have
$$e_i^1[b_i^*[\hat{e}_i]] = b_i^1(t_i^*[\hat{e}_i])(c_j) = \hat{e}_i^1(c_j)$$
as follows from condition (2) in definition III.9.

Induction Step: Assume that $e_i^k[b_i^*[\hat{e}_i]] = \hat{e}_i^k$ for all $\hat{e}_i \in E_i$, $k \geq 1$, and both players i. For $k+1$, take some player i and some $\hat{e}_i \in E_i$. We then have, for all $c_i^{k+1} \in C_i^{k+1}$,

$$\begin{aligned}
e_i^{k+1}[b_i^*[\hat{e}_i]](c_i^{k+1}) &= e_i^{k+1}[b_i^*[\hat{e}_i]](c_j, c_j^k) \\
&= \int_{\{c_j\} \times B_j} e_j^k[b_j](c_j^k) \, \mathrm{d}b_i^*[\hat{e}_i] \\
&= \hat{e}_i^1(c_j) \cdot e_j^k[b_j^*[e_j^*[c_j, \hat{e}_i]]](c_j^k) \\
&= \hat{e}_i^1(c_j) \cdot (e_j^*)^k[c_j, \hat{e}_i]](c_j^k) \\
&= \hat{e}_i^1(c_j) \frac{\hat{e}_i^{k+1}(c_j, c_j^k)}{\hat{e}_i^1(c_j)} \\
&= \hat{e}_i^{k+1}(c_i^{k+1})
\end{aligned}$$

where step 3 uses condition (2) in definition III.9 (note that $b_j^*[e_j^*[c_j, \hat{e}_i]]$ is the belief hierarchy induced by $t_j^*[e_j^*[c_j, \hat{e}_i]]$), step 4 uses the induction assumption, and step 5 uses condition (1) in definition III.9. This completes the induction. □

In what sense is $b_i^*[e_i]$ a "natural" representative of its equivalence class $[b_i^*[e_i]] = B_i(e_i)$?

One way in which $b_i^*[e_i]$ is natural is that it satisfies a **one-theory-per-choice** condition. That is, at each level of $b_i^*[e_i]$ condition (2) in definition III.9 associates exactly one type $t_j^*[e_j^*[e_i, c_j]]$ or, equivalently, one belief hierarchy $b_j^*[e_j^*[e_i, c_j]]$ with each opponent's choice c_j receiving positive probability. Note that, since the numbers $e_i^1(c_j)$ in condition (2) are uniquely identified, there is precisely one way of doing this. We can therefore think of $b_i^*[e_i]$ as the "sparsest" belief hierarchy that induces the expectation vector e_i. At each level of $b_i^*[e_i]$, the smallest amount of belief hierarchies necessary to induce e_i are invoked and every other belief hierarchy inducing e_i is less "sparse" in in that it requires strictly more types for being encoded.

A characteristic example of such a belief hierarchy is the belief encoded by your type t_y^1 in example III.1 above. Each choice that occurs at any levels of that types belief hierarchy is supported by exactly one type for you or Alice. This way, summarizing your belief hierarchy by its induced expectation vector comes without loss of information.

A less obvious reason why $b_i^*[e_i]$ is a natural representative of $B_i(e_i)$ is that it can be used as a reference belief hierarchy for characterizing belief in rationality. More precisely, we will now prove that if we want to check whether *some* belief hierarchy in an equivalence class $B_i(e_i)$ expresses up to k-fold or common belief in rationality, it is sufficient to check the canonical belief hierarchy $b_i^*[e_i]$.

Theorem III.11. *(Canonical Beliefs Give Rationality Its Best Shot)*
Let $e_i \in E_i$ be an expectation vector for player i. For any $k \geq 1$, there is a belief hierarchy b_i that induces e_i and expresses up to k-fold belief in rationality if, and only if, $b_i^*[e_i]$ does express up to k-fold belief in rationality.

Proof. The if-direction is clear by observation III.10. To prove the only-if-direction, we proceed by induction over $k \geq 1$.

Induction Start: Suppose that b_i expresses belief in the opponent's rationality. Then c_j is rational given b_j for every $(c_j, b_j) \in \text{Supp}(b_i)$. By definition of u_j, we therefore have

$$u_j(c_j, e_j[b_j]) \geq u_j(c'_j, e_j[b_j])$$

for all $c'_j \in C_j$ and every $(c_j, b_j) \in \text{Supp}(b_i)$.

Further, for all c_j such that $e_i^1(c_j) > 0$ and all measurable $A \subseteq C_j^\infty$, we have

$$e_j^*[e_i, c_j](A) = \frac{\int_{\text{Supp}(b_i)} 1_{\{c'_j = c_j\}} e_j[b_j](A)\, \mathrm{d}b_i}{e_i^1(c_j)}$$

as follows from condition (1) in definition III.9. Using belief linearity, this implies

$$u_j(c_j, e_j^*[e_i, c_j]) = \frac{\int_{\text{Supp}(b_i)} 1_{\{\hat{c}_j = c_j\}} u_j(c_j, e_j[b_j])\, \mathrm{d}b_i}{e_i^1(c_j)} \geq \frac{\int_{\text{Supp}(b_i)} 1_{\{\hat{c}_j = c'_j\}} u_j(c'_j, e_j[b_j])\, \mathrm{d}b_i}{e_i^1(c_j)} = u_j(c'_j, e_j^*[e_i, c_j])$$

for all $c'_j \in C_j$

By definition of u_i, every c_j with $e_i^1(c_j) > 0$ is then also rational given $b_j^*[e_j^*[e_i, c_j]]$, the belief hierarchy induced by $t_j^*[e_j^*[e_i, c_j]]$. Therefore c_j is rational given b_j for every $(c_j, b_j) \in \text{Supp}(b_i^*[e_i[b_i]])$ and, hence, $b_i^*[e_i[b_i]]$ expresses belief in the opponent's rationality.

Induction Step: Assume that, for all players i, if there is a belief hierarchy b_i that induces e_i and expresses up to k-fold belief in rationality, then $b_i^*[e_i]$ does express up to k-fold belief in rationality.

Now suppose that there is a b_i that induces e_i and expresses up to $(k+1)$-fold belief in rationality. Then, for every $(c_j, b_j) \in \text{Supp}(b_i)$, c_j is rational given b_j and b_j expresses up to k-fold belief in rationality. As before, for all c_j such that $e_i^1(c_j) > 0$, we have

$$e_j^*[e_i, c_j](A) = \frac{\int_{\text{Supp}(b_i)} 1_{\{c'_j = c_j\}} e_j[b_j](A)\, \mathrm{d}b_i}{e_i^1(c_j)} \tag{1}$$

for all measurable $A \subseteq C_j^\infty$ and, using belief linearity,

$$u_j(c_j, e_j^*[e_i, c_j]) = \frac{\int_{\text{Supp}(b_i)} 1_{\{\hat{c}_j = c_j\}} u_j(c_j, e_j[b_j])\, \mathrm{d}b_i}{e_i^1(c_j)} \geq \frac{\int_{\text{Supp}(b_i)} 1_{\{\hat{c}_j = c'_j\}} u_j(c'_j, e_j[b_j])\, \mathrm{d}b_i}{e_i^1(c_j)} = u_j(c'_j, e_j^*[e_i, c_j])$$

for all $c'_j \in C_j$.

Hence, every c_j with $e_i^1(c_j) > 0$ is rational given $b_j^*[e_j^*[e_i, c_j]]$ (the belief hierarchy induced by

$t_j^*[e_j^*[e_i, c_j]]$). Further, as $e_j[b_j]$ is linear in b_j, it follows from equation (1) that

$$e_j^*[e_i, c_j](A) = e_j\left[\frac{\int_{\text{Supp}(b_i)} 1_{\{c_j'=c_j\}} b_j \, \mathrm{d}b_i}{e_i^1(c_j)}\right](A)$$

for all measurable $A \subseteq C_j^\infty$.

Since

$$\tilde{b}_j = \frac{\int_{\text{Supp}(b_i)} 1_{\{c_j'=c_j\}} b_j \, \mathrm{d}b_i}{e_i^1(c_j)}$$

is a convex combination of belief hierarchies expressing up to k-fold belief in rationality, it expresses up to k-fold belief in rationality too. And since $\tilde{b}_j$ induces $e_j^*[e_i, c_j]$ it follows from the induction assumption that $b_j^*[e_j^*[e_i, c_j]]$ expresses up to k-fold belief in rationality.

Therefore c_j is rational under up to k-fold belief in rationality given b_j for every $(c_j, b_j) \in \text{Supp}(b_i^*[e_i])$ and, hence, $b_i^*[e_i]$ expresses up to $k+1$-fold belief in rationality.

The induction, and thereby the proof, are now complete. □

Given theorem III.11, it makes sense to define:

Definition III.12. *(Up to k-Fold and Common Belief in Rationality in Expectations)*
Let Γ be an expectation-based game and let e_i be an expectation vector for player i. e_i is said to **express up to k-fold (common) belief in rationality** *if $b_i^*[e_i]$ expresses up to k-fold (common) belief in rationality.*

Theorem III.11 ensures that, in expectation-based games, the expectational version of belief in rationality is necessary and sufficient for the fundamental version of belief in rationality in terms of belief hierarchies (see definition II.5). As we will see further below, this will allow us to characterize *common belief in rationality* in expectation-based games using an algorithm that only operates on expectation vectors, without any further recourse to belief hierarchies.

IV Possibility of Common Belief in Rationality

In this section, we explore the conditions under which *common belief in rationality* is possible in expectation-based games. For general static psychological games, Jagau and Perea (2018) present a condition called *preservation of rationality at infinity* which guarantees the existence of belief hierarchies expressing common belief in rationality. A stronger sufficient condition, *belief continuity*, was first introduced by Geanakoplos et al. (1989) and guarantees not only the possibility of common belief in rationality, but also the existence of *psychological Nash equilibrium*. Also, Jagau and Perea (2018) prove that belief continuity guarantees that the *belief-in-rationality operator* is

closed. That is, any choice that is rational under up to k-fold belief in rationality for every finite k is automatically rational under common belief in rationality in a belief-continuous game.

In the present section, we first show by means of an example that the class of expectation-based games also encompasses games that do *not preserve rationality at infinity* and for which common belief in rationality is impossible.

Subsequently, we prove that *belief continuity* takes a considerably simpler form once we restrict to expectation-based games. Intuitively, it turns out that continuity of the psychological Bernoulli utility functions v_i for all players i is necessary and sufficient for full-fledged belief continuity. One nice consequence of this is that any belief-finite expectation-based game has a closed *belief-in-rationality operator* and a *psychological Nash equilibrium*.[4]

A. Preservation of Rationality at Infinity

To begin, we recap preservation of rationality at infinity for general games.

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.

In their proof that definition IV.1 is sufficient for common belief in rationality, Jagau and Perea (2018) only use *probability-one belief hierarchies*, that is belief hierarchies where some choice is assigned probability one at every level of the belief. Going back to definition III.9, it is easy to check that every probability-one belief hierarchy b_i is an isolated canonical belief hierarchy corresponding to some extreme full expectation vector $c_i^\infty[b_i] \in C_i^\infty$. Theorem IV.2 in Jagau and Perea (2018) therefore directly translates into the following sufficient condition for possibility of common belief in rationality in expectation-based games:

Corollary IV.2. *(Possibility of Common Belief in Rationality in Expectation-Based Games)*
Let $\Gamma = (C_i, B_i, u_i)_{i \in I}$ be an expectation-based game and let $c_i \in C_i$ be a choice and $c_i^\infty \in C_i^\infty$ an extreme full expectation vector for some player $i \in I$ in it. Suppose that whenever, for every $n \geq 1$, there is some $\hat{c}_i^\infty \in C_i^\infty$ with $\hat{c}_i^n = c_i^n$ such that $v_i(c_i, \hat{c}_i^\infty) \geq v_i(c_i', \hat{c}_i^\infty)$ for all $c_i' \in C_i$, then also $v_i(c_i, c_i^\infty) \geq v_i(c_i', c_i^\infty)$, for all $c_i' \in C_i$. Then, for every player i, there exists an expectation vector $e_i \in E_i$ that expresses common belief in rationality.

As it turns out the class of expectation-based games also encompasses games that do *not* preserve rationality at infinity so that possibility of common belief in rationality is not guaranteed.

[4] Neither of these is true for general psychological games, see examples IV.9 and VI.9 in Jagau and Perea (2018).

For an example of an expectation-based game that does not preserve rationality at infinity and also does not allow for common belief in rationality, consider the following slight variation of the *modified bravery game* from Jagau and Perea (2018), example **IV.3**:

Example IV.3. (Common Belief in Rationality May Be Impossible in Expectation-Based Games)
Expectational Bravery Game
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*.

Let $c_1^\infty(timid) = (\star, timid, \star, timid, \star, timid, \dots)$ be the extreme vector of expectations 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 player 1 maximize psychological expected utility with a Bernoulli-utility function given by $v_1(timid, c_1^\infty(timid)) = 0$ and $v_1(bold, c_1^\infty(timid)) = 1$, whereas $v_1(timid, c_1^\infty) = 1$ and $v_1(bold, c_1^\infty) = 0$ for every other extreme vector of expectations $c_1^\infty \neq c_1^\infty(timid)$. Hence, choice *timid* is optimal for player 1 whenever $e_1(c_1^\infty(timid)) \leq \frac{1}{2}$ and *bold* is optimal for $e_1(c_1^\infty(timid)) \geq \frac{1}{2}$. The game is summarized in table 4.

Table 4: Expectational Bravery Game

	$c_1^\infty(timid)$	$c_1^\infty \neq c_1^\infty(timid)$
timid	0	1
bold	1	0

We now show that this game does not preserve rationality at infinity. To see this, for every even k, let $bold^k|c_1^\infty(timid)$ be the extreme-vector of expectations where the kth entry of $c_1^\infty(timid)$ is replaced by *bold*. Since $bold^k|c_1^\infty(timid) \neq c_1^\infty(timid)$, we have $v_1(timid, bold^k|c_1^\infty(timid)) > v_1(bold, bold^k|c_1^\infty(timid))$ for every $k \geq 1$. However $v_1(timid, c_1^\infty(timid)) < v_1(bold, c_1^\infty(timid))$, so that choice *timid* is not optimal for the extreme vector of expectations $c_1^\infty(timid)$.

Since preservation of rationality at infinity fails, it is not guaranteed that common belief in rationality is possible in this game. In appendix A we show that indeed there are no belief hierarchies for the given game that express common belief in rationality. The argument coincides almost word for word with the impossibility example IV.3 in Jagau and Perea (2018).

While preservation of rationality at infinity can therefore not be replaced by a weaker condition within the realm of expectation-based psychological games, it turns out that *belief continuity*, which guarantees existence of a *psychological Nash equilibrium* and closedness of the *belief-in-rationality operator*, can be considerably weakened.

B. Belief Continuity

We start by recapturing the general definition of belief continuity and related results.[5] 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 $C_{-i} \times B_{-i}^{k-1}$.

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 $k \in \mathbb{N}$ 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 $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$.

As shown in Geanakoplos et al. (1989) and Jagau and Perea (2018), belief continuity has two nice implications. The first one is that the belief-in-rationality operator is closed. So whenever a choice is rational under up to k-fold belief in rationality for every finite k, we can be sure that it is also rational under common belief in rationality.

Theorem IV.5. *(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. See Jagau and Perea (2018), theorem VI.8. □

The second one is that there always exists a psychological Nash equilibrium. That is, we can find a combination of *simple belief hierarchies* for all players that expresses common belief in rationality.[6]:

Definition IV.6. *(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 $c_i' \in C_i$.*

[5] Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009) define belief continuity as continuity of the full belief hierarchy given the product topology on B_i. In Jagau and Perea (2018), it is shown that that definition exactly coincides with the more explicit one given here.

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

Theorem IV.7. *(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. See Geanakoplos et al. (1989), theorem 1. □

We will now show that, for expectation-based games, belief continuity can be given a much simpler characterization in terms of a continuity restriction on psychological Bernoulli utilities v_i. Specifically, we will show that if the functions v_i are continuous in the product topology on C_i^∞, then belief linearity immediately yields full belief continuity of u_i.

As a preliminary, we show that the Lévy-Prokhorov distance between kth-order beliefs $b_i^k, \hat{b}_i^k \in B_i^k$ can be used to bound the Lévy-Prokhorov distance $d(e_i^k[b_i], e_i^k[\hat{b}_i])$ between the derived kth-order expectations $e_i^k[b_i], e_i^k[\hat{b}_i] \in E_i^k$.

Lemma IV.8. *(Distance between Beliefs Bounds Distance between Expectations)*

$$d(b_i^k, \hat{b}_i^k) \geq \frac{d(e_i^k[b_i], e_i^k[\hat{b}_i])}{2^{k-1}} \text{ for both players } i \text{ and } k \geq 1.$$

Proof.
To start, note that $d(e_i^k, \hat{e}_i^k) = \frac{1}{2}\sum_{c_i^k \in C_i^k} |e_i^k(c_i^k) - \hat{e}_i^k(c_i^k)|$ for all $k \geq 1$, which is a straightforward consequence of the fact that e_i^k has a discrete support.

Formally, consider the metric space $(C_i^k, \hat{d})$ with $\hat{d}$ the discrete metric. The Lévy-Prokhorov distance between k-th order expectations is given by

$$d(e_i^k, \hat{e}_i^k) = \inf\{\varepsilon > 0 |\ e_i^k(A) \leq \hat{e}_i^k(A^\varepsilon) + \varepsilon \text{ and } \hat{e}_i^k(A) \leq e_i^k(A^\varepsilon) + \varepsilon \text{ for all } A \in \mathcal{P}(C_i^k)\}$$

where $A^\varepsilon = \{c_i^k \in C_i^k |\ \exists \hat{c}_i^k \in A,\ \hat{d}(c_i^k, \hat{c}_i^k) < \varepsilon\} = \begin{cases} A, & \varepsilon < 1 \\ C_i^k, & \text{else} \end{cases}$. It follows that

$$\begin{aligned} d(e_i^k, \hat{e}_i^k) &= \inf\{\varepsilon > 0 |\ e_i^k(A) \leq \hat{e}_i^k(A) + \varepsilon \text{ and } \hat{e}_i^k(A) \leq e_i^k(A) + \varepsilon \text{ for all } A \in \mathcal{P}(C_i^k)\} \\ &= \max_{A \in \mathcal{P}(C_i^k)} |e_i^k(A) - \hat{e}_i^k(A)| \\ &= \sum_{c_i^k \in C_i^k} \max\{e_i^k(c_i^k) - \hat{e}_i^k(c_i^k), 0\} \\ &= \frac{1}{2}\sum_{c_i^k \in C_i^k} |e_i^k(c_i^k) - \hat{e}_i^k(c_i^k)|, \end{aligned}$$

as desired.
Given this, we can prove the theorem by induction over $k \geq 1$.
Induction Start: For $k = 1$, the result trivially follows from $b_i^1 = e_i^1$.

Induction Step: Now assume that $d(e_i^k[b_i], e_i^k[\hat{b}_i]) \leq 2^{k-1} d(b_i^k, \hat{b}_i^k)$ for both players i. Then, for $k+1$, we have

$$\begin{aligned}\frac{d(e_i^{k+1}[b_i], e_i^{k+1}[\hat{b}_i])}{2^{k-1}} &= \frac{1}{2^k} \sum_{c_i^{k+1} \in C_i^{k+1}} \left| e_i^{k+1}[b_i](c_i^{k+1}) - e_i^{k+1}[\hat{b}_i](c_i^{k+1}) \right| \\ &= \sum_{(c_j, c_j^k) \in C_i^{k+1}} \left| \int_{\{c_j\} \times B_j} \frac{e_j^k[b_j](c_j^k)}{2^k} \, \mathrm{d}(b_i - \hat{b}_i) \right|.\end{aligned}$$

Since $2d(e_i^k, \hat{e}_i^k) = \sum_{c_i^k \in C_i^k} |e_i^k(c_i^k) - \hat{e}_i^k(c_i^k)| \leq 2^k d(b_j^k, \hat{b}_j^k)$ by the induction assumption, the function $f(c_j', b_j^k) = 1_{\{c_j' = c_j\}} e_j^k[b_j](c_j^k)$ is Lipschitz-continuous with constant $Lip(f)$ at most 2^k on $C_j \times B_j$. It then follows from the dual characterization of the Wasserstein metric d_W in Kantorovich and Rubinstein (1958) that

$$\begin{aligned}\max_{(c_j, c_j^k) \in C_i^{k+1}} \left\{ \left| \int_{\{c_j\} \times B_j} \frac{e_j^k[b_j](c_j^k)}{2^k} \, \mathrm{d}(b_i - \hat{b}_i) \right| \right\} &\leq \sup_{\substack{f \text{ s.th.} \\ Lip(f) \leq 1}} \left\{ \left| \int_{C_j \times B_j} f(c_j, b_j^k) \, \mathrm{d}(b_i - \hat{b}_i) \right| \right\} \\ &= d_W(b_i^{k+1}, \hat{b}_i^{k+1}) \\ &\leq 2d(b_i^{k+1}, \hat{b}_i^{k+1})\end{aligned}$$

where the last inequality follows from Huber (1981), corollary 4.3. The induction is now complete. □

We are now ready to characterize belief continuity in expectation-based games.

Theorem IV.9. *(Belief Continuity in Expectation-Based Games)*
An expectation-based game is belief continuous if, and only if, for every player i, every choice c_i, every $\varepsilon > 0$, and every extreme expectation vector c_i^∞, there is some $k \geq 1$ such that for every extreme expectation vector $\hat{c}_i^\infty \in C_i^\infty$ with $c_i^k = \hat{c}_i^k$ we have that $|v_i(c_i, c_i^\infty) - v_i(c_i, \hat{c}_i^\infty)| < \varepsilon$.

Proof.

⇒ To see that belief continuity in expectation-based games implies the condition in the theorem, consider some c_i^∞ and the corresponding probability-one belief hierarchy b_i. By belief-continuity, there exist $k \in \mathbb{N}$ and $\delta > 0$ such that for any belief hierarchy $\hat{b}_i \in B_i$ with $d(b_i^m, \hat{b}_i^m) < \delta$ for all $m \leq k$, we have $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| < \varepsilon$. Now pick $\hat{c}_i^\infty$ such that $c_i^k = \hat{c}_i^k$ and let $\hat{b}_i$ be the belief hierarchy that induces $\hat{c}_i^\infty$. Since b_i and $\hat{b}_i$ are probability-one belief hierarchies, we also have $b_i^k = \hat{b}_i^k$ and therefore $d(b_i^m, \hat{b}_i^m) = 0$ for all $m \leq k$. Hence $|u_i(c_i, b_i) - u_i(c_i, \hat{b}_i)| = |v_i(c_i, c_i^\infty) - v_i(c_i, \hat{c}_i^\infty)| < \varepsilon$, as desired.

⇐ To prove that also the reverse is true, consider an expectation-based game where psychological Bernoulli utility functions v_i satisfy the above condition for all players i and let c_i, b_i, and

$\varepsilon > 0$ be given. Also, let $M > |v_i(c_i, c_i^\infty)|$ for all $c_i^\infty \in C_i^\infty$.[7]

Now choose $k \geq 1$ such that $|v_i(c_i, c_i^\infty) - v_i(c_i, \hat{c}_i^\infty)| < \frac{\varepsilon}{3}$ for all $c_i^\infty, \hat{c}_i^\infty \in C_i^\infty$ with $c_i^k = \hat{c}_i^k$. This is possible since the set C_i^k of extreme kth-order expectations is a finite set. Moreover, fix an arbitrary $\tilde{c}_i^\infty \in C_i^\infty$ and, for all $c_i^k \in C_i^k$, let $c_i^k|\tilde{c}_i^\infty$ be the extreme full expectation vector where the first k entries of $\tilde{c}_i^\infty$ are replaced by c_i^k.

Next, choose $\hat{b}_i$ such that $d(b_i^m, \hat{b}_i^m) < \frac{\varepsilon}{3(2^k)M}$ for all $m \leq k$ and let $e_i = e_i[b_i]$ and $\hat{e}_i = e_i[\hat{b}_i]$ denote the corresponding expectation vectors. Then, using the representation from theorem III.5,

$$
\begin{aligned}
|u_i(c_i, e_i) - u_i(c_i, \hat{e}_i)| &= \left| \int_{C_i^\infty} v_i(c_i, c_i^\infty) \, \mathrm{d}e_i - \int_{C_i^\infty} v_i(c_i, c_i^\infty) \, \mathrm{d}\hat{e}_i \right| \\
&= \left| \sum_{c_i^k \in C_i^k} e_i^k(c_i^k) v_i(c_i, c_i^k|\tilde{c}_i^\infty) - \left(\sum_{c_i^k \in C_i^k} e_i^k(c_i^k) v_i(c_i, c_i^k|\tilde{c}_i^\infty) - \int_{C_i^\infty} v_i(c_i, c_i^\infty) \, \mathrm{d}e_i \right) \right. \\
&\quad \left. - \left(\sum_{c_i^k \in C_i^k} \hat{e}_i^k(c_i^k) v_i(c_i, c_i^k|\tilde{c}_i^\infty) - \left(\sum_{c_i^k \in C_i^k} \hat{e}_i^k(c_i^k) v_i(c_i, c_i^k|\tilde{c}_i^\infty) - \int_{C_i^\infty} v_i(c_i, c_i^\infty) \, \mathrm{d}\hat{e}_i \right) \right) \right| \\
&< \left| \sum_{c_i^k \in C_i^k} e_i^k(c_i^k) v_i(c_i, c_i^k|\tilde{c}_i^\infty) - \sum_{c_i^k \in C_i^k} \hat{e}_i^k(c_i^k) v_i(c_i, c_i^k|\tilde{c}_i^\infty) \right| + \frac{2\varepsilon}{3} \\
&\leq \sum_{c_i^k \in C_i^k} \left| e_i^k(c_i^k) - \hat{e}_i^k(c_i^k) \right| \left| v_i(c_i, c_i^k|\tilde{c}_i^\infty) \right| + \frac{2\varepsilon}{3} \\
&\leq \sum_{c_i^k \in C_i^k} \left| e_i^k(c_i^k) - \hat{e}_i^k(c_i^k) \right| M + \frac{2\varepsilon}{3} \\
&= 2Md(e_i^k, \hat{e}_i^k) + \frac{2\varepsilon}{3} \\
&\leq 2^k M d(b_i^k, \hat{b}_i^k) + \frac{2\varepsilon}{3} < \varepsilon.
\end{aligned}
$$

Note that for the sixth step we used $d(e_i^k, \hat{e}_i^k) = \frac{1}{2} \sum_{c_i^k \in C_i^k} |e_i^k(c_i^k) - \hat{e}_i^k(c_i^k)|$, which is a straightforward consequence of the fact that e_i^k has a discrete support. The seventh step then follows from lemma IV.8.

□

Based on theorem IV.9, it is not hard to see that every belief-finite, expectation-based game is belief continuous so that all the nice conditions for belief-continuous games already hold whenever utilities in a game only depend on finite-orders of expectations.

[7] Here we use our initial assumption that u_i is bounded (see definition II.1).

Corollary IV.10. *(Existence in Belief-Finite Expectation-Based Games)*
If an expectation-based game is belief-finite, then

1. *there exists a pair of expectations vectors for both players that express common belief in rationality,*
2. *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,*
3. *there exists a psychological Nash equilibrium.*

V Common Belief in Rationality Characterized

In this subsection we develop an elimination procedure for two-player expectation-based games, called *iterated elimination of choices and expectations*, which is similar to *iterated elimination of choices and belief hierarchies* as introduced in Jagau and Perea (2018). The only difference is that, instead of belief hierarchies, we will keep track of vectors of expectations $e_i[b_i]$ expressing increasing level of belief in rationality.

As we saw in theorem III.12, it makes sense to think of an expectation vector as expressing up to k-fold or common belief in rationality to the extent that the canonical belief hierarchy $b_i^*[e_i]$ associated with each class of expectationally equivalent belief hierarchies $B_i(e_i)$ necessarily expresses up to k-fold or common belief in rationality if any belief hierarchy in $B_i(e_i)$ does.

We first introduce some additional notation that is needed for the new procedures. For every choice $c_j \in C_j$, we denote by $[c_j]$ the degenerate probability distribution on C_j that assigns probability 1 to c_j. Let $[C_j] := \{[c_j] | c_j \in C_j\}$ be the set of all such degenerate probability distributions. Then, as a consequence of Theorem III.3, every $e_i \in E_i$ can be identified with a vector in $\text{Conv}([C_j] \times E_j)$ where Conv stands for "convex hull".

We are now ready to define *iterated elimination of choices and expectations*:

Procedure V.1. *(Iterated Elimination of Choices and Expectations)*

Step 1: For both players i define

$$R_i(1) := \{([c_i], e_i) \in [C_i] \times E_i \quad | \quad u_i(c_i, e_i) \geq u_i(c_i', e_i), \quad \forall c_i' \in C_i\}.$$

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

$$R_i(k) := \{([c_i], e_i) \in R_i(k-1) \quad | e_i \in \text{Conv}(R_j(k-1))\}.$$

We finally define, for both players i,

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

In the following theorem we show that, for every two-person expectation based game, this procedure selects exactly those combinations of choices and expectations that are possible under common belief in rationality.

Theorem V.2. *(The Algorithm Works)*
Take a two-player expectation-based psychological game Γ.

1. *For every k, the expectation vectors e_i that express up to k-fold belief in rationality are exactly the expectation vectors in $\text{proj}_{E_i}(R_i(k+1))$ surviving $k+1$ consecutive steps of elimination of choices and expectations. Also the choices that can be made under up to k-fold belief in rationality are exactly the choices in the projection $\text{proj}_{C_i}(R_i(k+1))$.*

2. *The expectation vectors e_i that express common belief in rationality, if existent, are exactly the expectation vectors in $\text{proj}_{E_i}(R_i(\infty))$ that survive iterated elimination of choices and expectations. The choices that can be rationally made under common belief in rationality are exactly the choices in the projection $\text{proj}_{C_i}(R_i(\infty))$.*

Proof.
It suffices to prove part 1 of the theorem. Part 2 then follows from the fact that, whenever $([c_i], e_i) \in R_i(\infty)$, there is a uniquely identified belief hierarchy, namely $b_i^*[e_i]$, that induces e_i and expresses up to k-fold belief in rationality for any $k \geq 1$ and hence common belief in rationality.

$\Rightarrow$ We start by showing that we have $([c_i], e_i) \in R_i(k+1)$ for every $(c_i, e_i) \in C_i \times E_i$ such that e_i expresses up to k-fold belief in rationality and rationalizes c_i. We proceed by induction over $k \geq 0$.

Induction start: For $k = 0$, the statement is true by definition of $R_i(1)$.

Induction step: Assume that, for all players i, $([c_i], e_i) \in R_i(k+1)$ whenever e_i expresses up to k-fold belief in rationality and rationalizes c_i. Now let e_i express up to $k+1$-fold belief in rationality and rationalize c_i. We must show that $([c_i], e_i) \in R_i(k+2)$.

As e_i expresses up to $k+1$-fold belief in rationality, it follows in particular that e_i expresses up to k-fold belief in rationality. Hence, by the induction assumption, $([c_i], e_i) \in R_i(k+1)$.

Since e_i expresses up to $k+1$-fold belief in rationality, $b_i^*[e_i]$ expresses up to $k+1$-fold belief in rationality. So all $(c_j, b_j^*[e_j^*[e_i, c_j]]) \in \text{Supp}(b_i^*[e_i])$ are such that $b_j^*[e_j^*[e_i, c_j]]$ rationalizes c_j under up to k-fold belief in rationality. Hence, all $e_j^*[e_i, c_j]$ express up to k-fold belief in rationality and, therefore, using the induction assumption, $([c_j], e_j^*[e_i, c_j]) \in R_j(k+1)$.

Hence, we have

$$e_i = \sum_{c_j : e_i^1(c_j) > 0} e_i^1(c_j)([c_j], e_j^*[e_i, c_j]) \in \text{Conv}(R_j(k+1))$$

and, consequently, $([c_i], e_i) \in R_i(k+2)$, such that the first direction is established.

$\Leftarrow$ For the reverse direction we show, by induction over $k \geq 0$, that e_i expresses up to k-fold belief in rationality and rationalizes c_i for every $([c_i], e_i) \in R_i(k+1)$.

Induction start: For $k = 0$, the first part follows by definition of $R_i(1)$.
Induction step: Assume that, for all players i, e_i expresses up to k-fold belief in rationality and rationalizes c_i whenever $([c_i], e_i) \in R_i(k+1)$.

Now take some $([c_i], e_i) \in R_i(k+2)$. Then $e_i \in \text{Conv}(R_j(k+1))$. Since, by definition III.9,

$$e_i = \sum_{c_j : e_i^1(c_j)} e_i^1(c_j) \cdot ([c_j], e_j^*[e_i, c_j]),$$

this implies $([c_j], e_j^*[e_i, c_j]) \in R_j(k+1)$ for every c_j with $e_i^1(c_j) > 0$. Using the linearity of u_i and the induction assumption $e_j^*[e_i, c_j]$ then rationalizes c_j under up to k-fold belief in rationality so that, in particular, $b_j^*[e_j^*[e_i, c_j]]$ expresses up to k-fold belief in rationality for every c_j with $e_i^1(c_j) > 0$. By construction, it follows that $b_i^*[e_i]$ assigns full probability to opponent's choices and belief hierarchies $(c_j, b_j^*[e_j^*[e_i, c_j]])$ such that $b_j^*[e_j^*[e_i, c_j]]$ rationalizes c_j under up to k-fold belief in rationality. Hence, $b_i^*[e_i]$ expresses up to $k+1$-fold belief in rationality and rationalizes c_i, which establishes the reverse direction.

□

As theorem V.2 shows, even in the most general expectation-based games, common belief in rationality can be analyzed without any further recourse to belief hierarchies. So expectation vectors, in a way, can be used as the primitives that define a state of the world from the point of view of both players i.

Similar to our result from theorem **VI.5** in Jagau and Perea (2018) for general belief-finite games, belief-finite expectation-based games allow for characterizing common belief in rationality with a yet simpler procedure, *iterated elimination of choices and nth-order expectations. Iterated elimination of choices and nth-order expectations* amounts to a massive computational simplification in the analysis of expectation-based games: As we will see below, that procedure can be implemented as a sequence of linear programs.

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

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

Our procedure proceeds by iteratively eliminating choices and vectors of up to nth-order expectations.

Procedure V.3. *(Iterated Elimination of Choices and nth-Order Expectations)*

Step 1*: For both players i define*

$$R_i^n(1) := \{([c_i], e_i^n) \in [C_i] \times E_i^n \quad | \quad \exists e_i^{n+1} \in E_i^{n+1} \quad with \quad proj_{E_i^n} e_i^{n+1} = e_i^n \quad such\ that \quad u_i(c_i, e_i^{n+1}) \geq u_i(c_i', e_i^{n+1}) \quad \forall c_i' \in C_i\}.$$

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

$$R_i^n(k) := \{([c_i], e_i^n) \in R_i^n(k-1) \quad | \quad \exists e_i^{n+1} \in \mathrm{Conv}(R_j^n(k-1)) \quad with \quad proj_{E_i^n} e_i^{n+1} = e_i^n \quad such\ that \quad u_i(c_i, e_i^{n+1}) \geq u_i(c_i', e_i^{n+1}) \quad \forall c_i' \in C_i\}.$$

We finally define, for both players i,

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

Clearly, E_i^n is a convex set with finitely many extreme points and every $e_i^n \in E_i^n$ can be identified with a vector in $\mathrm{Conv}([C_j] \times E_j^{n-1})$.

Iterated elimination of choices and nth-order expectations provides a straightforward generalization of the characterization of common belief in rationality in traditional games: Whenever utility depends on at most $n+1$th-order expectations, we need to track choices and nth-order expectations. So, in particular, when utility depends only on first-order expectations or, equivalently, if we are dealing with a traditional games, we can resort to the familiar procedure *iterated elimination of strictly dominated choices.*

To state the characterization theorem, it will be useful to define:

Definition V.4. *(Consistency with up to k-Fold and Common Belief in Rationality)*
A combination $(c_i, e_i^n) \in C_i \times E_i^n$ of choices and nth-order expectations for player i is

a) **consistent with up to k-fold belief in rationality** *if there exists a full expectation vector $e_i \in E_i$ that expresses up to k-fold belief in rationality, induces e_i^n, and rationalizes c_i.*

b) **consistent with common belief in rationality** *if there exists a full expectation vector $e_i \in E_i$ that expresses common belief in rationality, induces e_i^n, and rationalizes c_i.*

We are now ready to state theorem V.5:

Theorem V.5. *(The Algorithm Works)*
Take a two-player belief-finite psychological game Γ *that is expectation-based and where utility functions depend only on* $n+1$*th- order expectations.*

1. *For all* $k \geq 0$*, a choice-expectation combination* $(c_i, e_i^n) \in C_i \times E_i^n$ *is consistent with up to* k*-fold belief in rationality if, and only if,* $([c_i], e_i^n) \in R_i^n(k+1)$.

2. *A choice-expectation combination* $(c_i, e_i^n) \in C_i \times E_i^n$ *is consistent with common belief in rationality if, and only if,* $([c_i], e_i^n) \in R_i^n(\infty)$.

Proof.
Part 1:

$\Rightarrow$ We start by showing that for every $(c_i, e_i^n) \in C_i \times E_i^n$ that is consistent with up to k-fold belief in rationality, we have that $([c_i], e_i^n) \in R_i^n(k+1)$. We proceed by induction over $k \geq 0$.

Induction start: For $k = 0$, take some $(c_i, e_i^n) \in C_i \times E_i^n$ that is consistent with up to 0-fold belief in rationality. Hence, c_i is optimal for some expectation vector $e_i \in E_i$ that induces e_i^n. Let e_i^{n+1} be the $n+1$th-order expectation induced by e_i. Then, $u_i(c_i, e_i^{n+1}) \geq u_i(c_i', e_i^{n+1})$ for all $c_i' \in C_i$ and $\text{proj}_{E_i^n} e_i^{n+1} = e_i^n$, which implies $([c_i], e_i^n) \in R_i^n(1)$.

Induction step: Assume that, for all players i, $([c_i], e_i^n) \in R_i^n(k+1)$ whenever (c_i, e_i^n) is consistent with up to k-fold belief in rationality. Let (c_i, e_i^n) be consistent with up to $k+1$-fold belief in rationality. We must show that $([c_i], e_i^n) \in R_i^n(k+2)$.

Since (c_i, e_i^n) is consistent with up to $k+1$-fold belief in rationality, it follows in particular that (c_i, e_i^n) is consistent with up to k-fold belief in rationality. Hence, by the induction assumption, $([c_i], e_i^n) \in R_i^n(k+1)$.

Since (c_i, e_i^n) is consistent with up to $k+1$-fold belief in rationality, we know that c_i is optimal for an expectation vector e_i that expresses up to $k+1$-fold belief in rationality, and that induces e_i^n. As we saw in the the proof of theorem V.2, e_i can then be written as a linear combination of opponent's choice expectation combinations $([c_j], e_j^*[e_i, c_j])$ where $e_j^*[e_i, c_j]$ rationalizes c_j under up to k-fold belief in rationality. By definition, $([c_j], (e_j^*)^n[e_i, c_j])$ must then be consistent with up to k-fold belief in rationality and, hence, $([c_j], (e_j^*)^n[e_i, c_j]) \in R_j^n(k+1)$. Clearly, it must also be possible to write e_i^{n+1} as a linear combination of the $([c_j], (e_j^*)^n[e_i, c_j])$ such that $e_i^{n+1} \in \text{Conv}(R_j^n(k+1))$ and hence, since e_i^{n+1} induces e_i^n, $([c_i], e_i^n) \in R_i^n(k+2)$. This establishes the first direction.

$\Leftarrow$ We now show that for every $([c_i], e_i^n) \in R_i^n(k+1)$, we have that (c_i, e_i^n) is consistent with up to k-fold belief in rationality. We again proceed by induction over $k \geq 0$.

Induction start: For $k = 0$, take some $([c_i], e_i^n) \in R_i^n(1)$. Then, by construction, there is some $e_i^{n+1} \in E_i^{n+1}$ with $\text{proj}_{E_i^n} e_i^{n+1} = e_i^n$ such that $u_i(c_i, e_i^{n+1}) \geq u_i(c_i', e_i^{n+1})$ for all $c_i' \in C_i$. Take an

arbitrary expectation vector $e_i \in E_i$ that induces e_i^{n+1}. Then, e_i induces e_i^n and rationalizes c_i. Hence, it follows that (c_i, e_i^n) is consistent with up to 0-fold belief in rationality.

Induction step: Assume that, for all players i, (c_i, e_i^n) is consistent with up to k-fold belief in rationality whenever $([c_i], e_i^n) \in R_i^n(k+1)$. Take some $([c_i], e_i^n) \in R_i^n(k+2)$. Then, by definition, there is some $e_i^{n+1} \in \text{Conv}(R_j^n(k+1))$ such that $\text{proj}_{E_i^n} e_i^{n+1} = e_i^n$ and $u_i(c_i, e_i^{n+1}) \geq u_i(c_i', e_i^{n+1})$ for all $c_i' \in C_i$. Hence, e_i^{n+1} can be written as a convex combination

$$e_i^{n+1} = \sum_{c_j \in C_j} \lambda(c_j) \cdot ([c_j], e_j^n[c_j])$$

where $\lambda(c_j) \geq 0$ for all $c_j \in C_j$, $\sum_{c_j \in C_j} \lambda(c_j) = 1$, and $([c_j], e_j^n[c_j]) \in R_j^n(k+1)$ for all $c_j \in C_j$ with $\lambda(c_j) > 0$.

By the induction assumption we know that every $([c_j], e_j^n[c_j]) \in R_j^n(k+1)$ is consistent with up to k-fold belief in rationality. Hence, for every $e_j^n[c_j]$, there is a full expectation vector $e_j[c_j]$ that induces $e_j^n[c_j]$, expresses up to k-fold belief in rationality, and rationalizes c_j. It follows that $([c_j], e_j) \in R_j(k+1)$ (with $R_j(k+1)$ as defined in procedure V.1) and, furthermore, that $e_i = \sum_{c_j \in C_j} \lambda(c_j) \cdot ([c_j], e_j[c_j]) \in \text{Conv}(R_j(k+1))$. Hence $([c_i], e_i) \in R_i^n(k+2)$. By theorem V.2, e_i therefore expresses up to $k+1$-fold belief in rationality. Since, by construction, e_i also induces e_i^n and rationalizes c_i, this implies that $([c_i], e_i^n)$ is consistent with up to $k+1$-fold belief in rationality. The second direction, and therefore part 1 of the proof, are now complete.

Part 2:

To see that any (c_i, e_i^n) that is consistent with common belief in rationality is such that $([c_i], e_i^n) \in R_i^n(\infty)$, just note that any choice-expectation combination that is consistent with common belief in rationality is automatically consistent with up to k-fold belief in rationality for any $k \geq 0$ and hence $([c_i], e_i^n) \in \bigcap_{k \geq 1} R_i^n(k) = R_i^n(\infty)$.

For the reverse direction, we vary a proof strategy from theorem VI.5 in Jagau and Perea (2018): Let $([c_i], e_i^n) \in R_i^n(\infty)$. Then, by part 1, there exists a sequence of expectation vectors $(e_i(k))_{k \in \mathbb{N}}$, where each $e_i(k)$ induces e_i^n and expresses up to k-fold belief in rationality. By theorem III.11, each $e_i(k)$ can be replaced with the corresponding canonical belief hierarchy $b_i^*[e_i(k)]$, yielding a sequence $(b_i^*[e_i(k)])_{k \in \mathbb{N}}$ of belief hierarchies where each $b_i^*[e_i(k)]$ induces e_i^n and expresses up to k-fold belief in rationality. Since B_i is Polish and thereby sequentially compact, $(b_i^*[e_i(k)])_{k \in \mathbb{N}}$ has a converging subsequence $(b_i^*[e_i'(k)])_{k \in \mathbb{N}}$, the limit of which we denote by $b_i^*[e_i'(\infty)]$. Clearly, also $b_i^*[e_i'(\infty)]$ induces e_i^n. Now, as shown by Jagau and Perea (2018) in their theorem VI.8, the sets $B_i(k, c_i)$ of belief hierarchies that rationalize c_i under up to k-fold belief in rationality are compact for every $k \geq 1$ in every belief-continuous game and, hence, in any belief-finite expectation-based game. So fix any k. Then $b_i^*[e_i'(m)] \in B_i(k, c_i)$ for all $m \geq k$ and, in particular, $b_i^*[e_i'(\infty)] \in B_i(k, c_i)$ by compactness of $B_i(k, c_i)$. Since k was arbitrarily chosen, it follows that $b_i^*[e_i'(\infty)] \in B_i(\infty, c_i) =$

$\bigcap_{k\geq 1} B_i(k, c_i)$. So $b_i^*[e_i'(\infty)]$ induces e_i^n and expresses common belief in rationality, which makes (c_i, e_i^n) consistent with common belief in rationality.

The proof of theorem V.5 is now complete. □

Theorem V.5 has a nice computational implication. To see this, note that, for both players i,

- $\text{Conv}([C_i] \times E_i^n)$ is a convex set with finitely many ($|C_i \times C_i^n|$) extreme points,
- $\text{Conv}(R_i^n(1))$ follows from imposing finitely many linear restrictions on $\text{Conv}([C_i] \times E_i^n)$.

Combining these two facts, we conclude that any $\text{Conv}(R_i^n(k))$, $k \geq 1$ is again a convex set with finitely many extreme points. We can therefore establish

Observation V.6. *(Linearity of Elimination of Choices and nth-Order Expectations)*
Iterated elimination of choices and nth-order expectations is a sequence of linear programs.

Given observation V.6, it is reasonable to ask whether the other nice property of traditional *iterated elimination of strictly dominated choices*, its finiteness, will also carry over. As we will show below, the answer, unfortunately, is no. The example presented below shows that, even in additive games, where iterated elimination of choices and up to nth-order expectations is a linear program, and already in the simplest non-degenerate case where both players only care about the first- and second-order expectation, the procedure does not necessarily terminate within finitely many steps.[8]

Example V.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.

[8] This does not mean that we have to give up on finiteness for *all* applications of expectation-based games. An important "degenerate" case is explored in Jagau and Perea (2018): In *unilateral psychological games*, one player's utility function depends on second-order beliefs while all other players' preferences depend on first-order beliefs. For this specific class of psychological games, it turns out that we can find a finite algorithm that characterizes common belief in rationality under assumptions that are even slightly more general than those of definition III.4.

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 expectation-based game with player set $I = \{y, a\}$ and choice sets $C_y = C_a = \{stay,\ run\}$. To write utility function down concisely, we define $m_i^2(c_i) = \int_{C_j \times B_j} b_j^1(c_i)\ \mathrm{d}b_i = \sum_{c_j} e_i^2(c_j, c_i)$ for $i \in \{a, y\}$. This expression, which we refer to as the *marginal 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, independent of i's first-order expectation.

Let your utility function now be given by

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

Similarly, Alice's utility function is given by

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

Similar to the above characterization result following observation III.7, we can represent Alice's and your preferences by *two pairs* of finite matrices containing the utilities that you and Alice derive from your extreme first-order expectations and your extreme marginal second-order expectations. This is shown in table 5 below.

Table 5: The Nightly Encounter

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

$+$

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

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

$+$

	m_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 expectation e_y^1 is *stay* and your marginal second-order expectation m_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 expectations]* 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 expectation, your marginal second-order expectation has to put at least this minimum probability on Alice expecting you to choose *run*. 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 marginal 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 marginal 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 expectations]* 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 expectations* to determine the combinations of choices and first-order expectations for you and Alice that are consistent with common belief in rationality. Since utility functions here depend linearly on first-order expectations and marginal second-order expectations, we can conveniently capture elimination steps as linear restrictions on the product space of first-order expectations and marginal second-order expectations for both you and Alice. To determine the set $R_y^1(1)$ of rational pairs of choices and first-order expectations for you, we first depict the pairs (e_y^1, m_y^2) of first-order expectations and marginal second-order expectations for which *stay* is rational, and the pairs for which *run* is rational. See the left-hand picture in Figure 2. Note that *stay* can only be rational for a pair of expectations (e_y^1, m_y^2) if $e_y^1(run) \leq \frac{4}{5}$. On the other hand, every first-order expectation e_y^1 can be extended to a pair (e_y^1, m_y^2) for which *run* is rational.

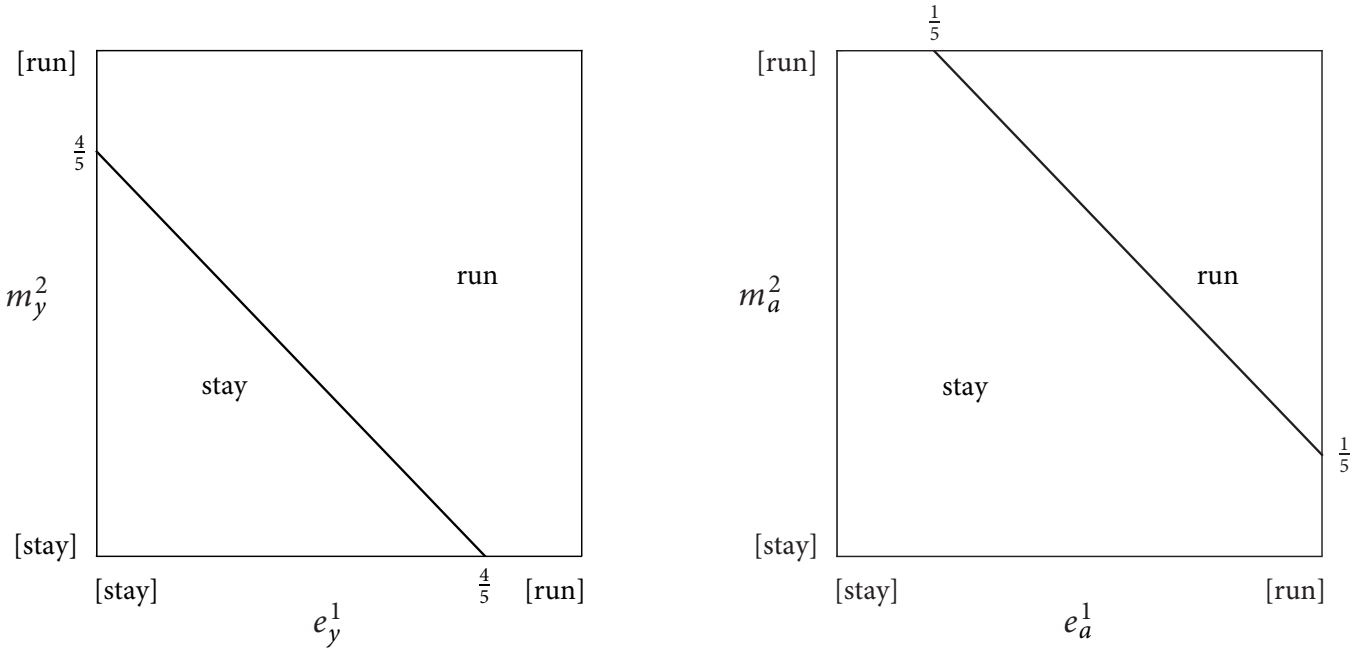
Hence, we conclude that

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

In a similar way we can derive $R_a^1(1)$ from the right-hand picture of Figure 2 and conclude that

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

Figure 2: Expectations for which Choices are Rational



The set of expectation combinations $(e_y^1, m_y^2))$ 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 3. Note that *stay* can only be rational for you for a pair of expectations (e_y^1, m_y^2) in $\text{Conv}(R_a^1(1))$ if $e_y^1(run) \leq \frac{2}{3}$. On the other hand, every first-order expectation e_y^1 can be extended to a pair (e_y^1, m_y^2) in $\text{Conv}(R_a^1(1))$ for which *run* is rational. Hence, we obtain that

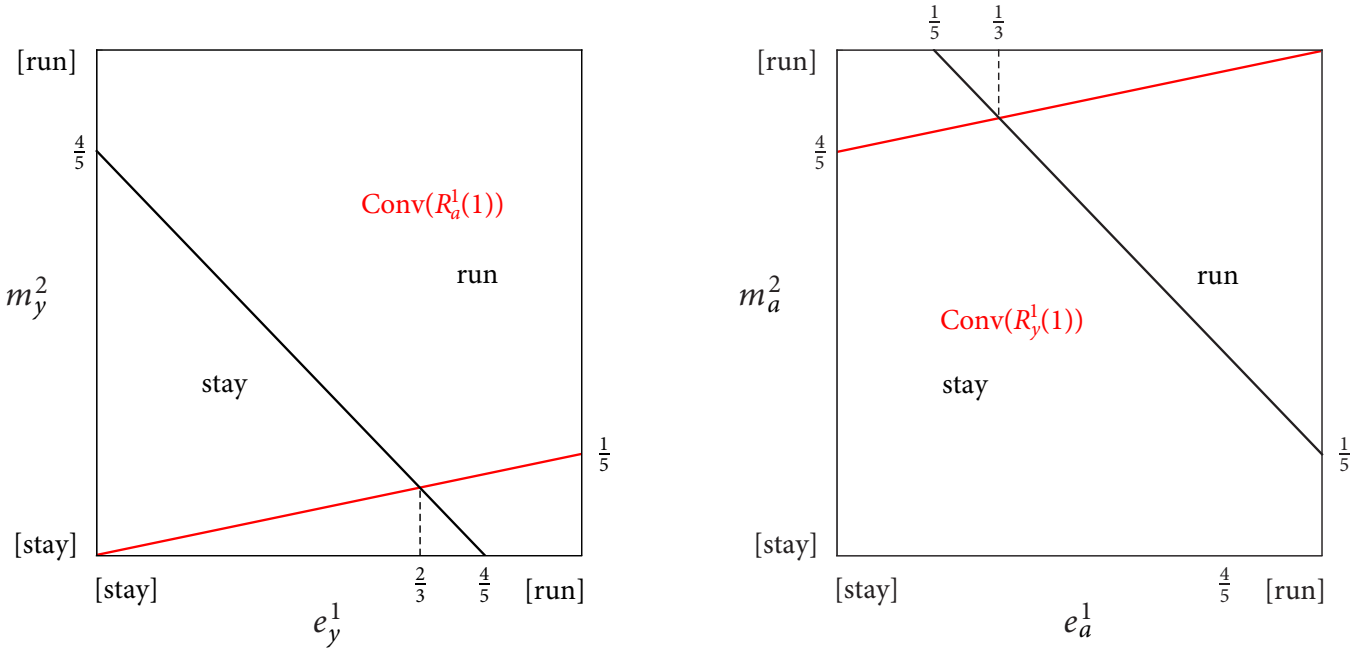
$$R_y^1(2) = \left\{(stay, e_y^1) | e_y^1(run) \leq \frac{2}{3}\right\} \quad \cup \quad \{(run, e_y^1) | e_y^1 \in \Delta(\{stay, run\})\}.$$

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 3. In the same way as above, we can derive from the right-hand picture of Figure 3 that

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

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 expectations* does not terminate within finitely many steps.

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



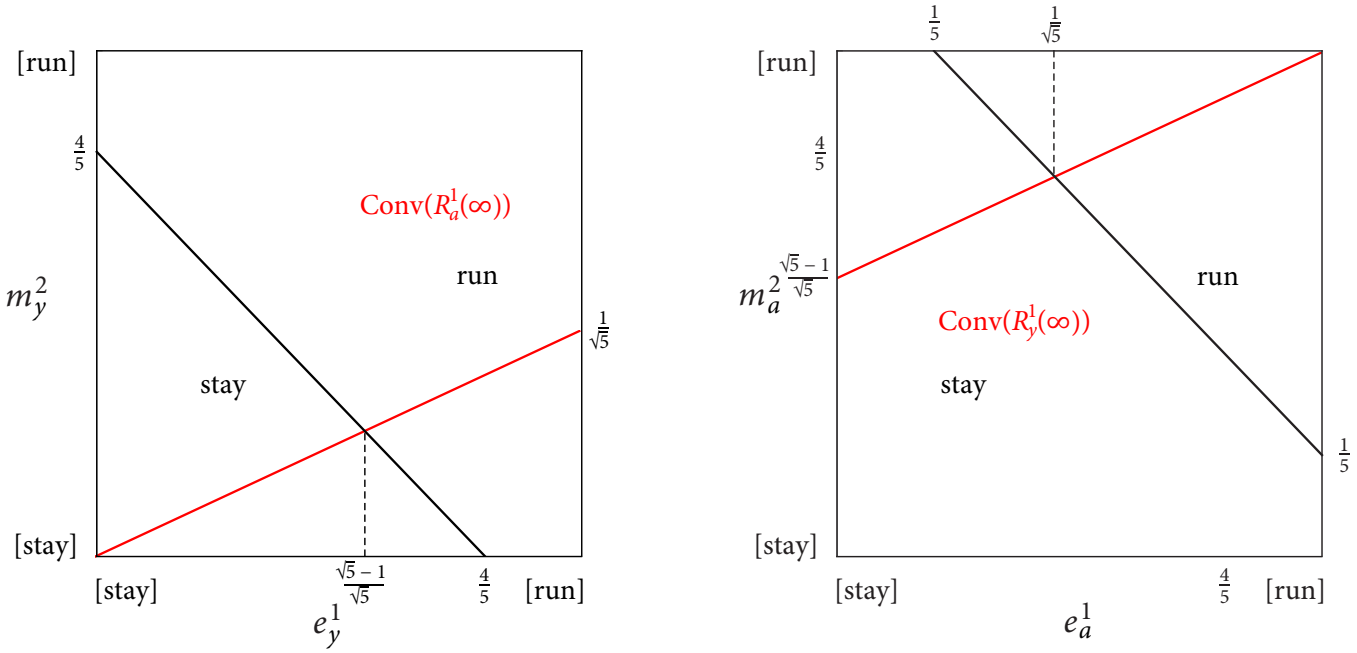
Finally, it can be verified that

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

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

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 4 shows how the sets $R_y^1(\infty)$ and $R_a^1(\infty)$ can be graphically constructed.

Figure 4: Convex hull of $R_a^1(\infty)$ and $R_y^1(\infty)$



VI Conclusion

Since the seminal paper by Geanakoplos et al. (1989), psychological game theory has become a widely applied theoretical framework to capture numerous belief-dependent motivations and emotional mechanisms in a natural and mathematically quantifiable way. Remarkably, however, there is a noticeable gap between the level of complexity at which existing theoretical contributions (cf. Geanakoplos et al. 1989, Kolpin 1992, Battigalli and Dufwenberg 2009, Jagau and Perea 2018) operate and the complexity of psychological games as studied in applied work. In particular, the highly non-linear structure of belief hierarchies prevents many properties we take for granted in traditional games from carrying over to general psychological games. Therefore, the question what

psychological games are nice to work with and what level of complexity we expect to deal with when analyzing them becomes an important issue in itself to investigate. Here, we have shrunk the gap between theory and applications by providing a first systematic extension of expected utility to psychological games and studying its properties. As we saw, *psychological expected utility* amounts to rather mild and intuitive restrictions on utility functions while at the same time rendering the analysis of the corresponding *expectation-based psychological games* massively simpler in computational terms. In particular, we could characterize common belief in rationality with a linear algorithm, *Iterated Elimination of Choices and Expectations*. Closely relating to this result, we could also extend the familiar matrix representation of traditional games to expectation-based games. While the procedure was found to sometimes require infinitely many steps to characterize common belief in rationality, its linearity ensures that any output of interest can easily be computed analytically or numerically, such that the added level of complexity of *Iterated Elimination of Choices and Expectations* relative to traditional *Iterated Elimination of Strictly Dominated Choices* remains pleasantly small.

Appendix

A Impossibility of Common Belief in Rationality in Example IV.3

Table 6: Expectational Bravery Game

	$c_1^\infty(timid)$	$c_1^\infty \neq c_1^\infty(timid)$
timid	0	1
bold	1	0

Here we prove that there is no belief hierarchy for player 1 that expresses common belief in rationality in the *Expectational Bravery Game*. We first show that the belief hierarchy inducing $c_1^\infty(timid)$ does not express common belief in rationality. By definition, $c_1^\infty(timid)$ is such that player 1 believes that player 2 believes that player 1 chooses *timid* given expectation $c_1^\infty(timid)$. However, *timid* is not optimal for $c_1^\infty(timid)$, and hence under $c_1^\infty(timid)$, player 1 believes that player 2 believes that player 1 chooses irrationally. It follows that the belief hierarchy inducing $c_1^\infty(timid)$ does not express up to 2-fold belief in rationality and, a fortiori, 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 inducing $c_1^\infty(timid)$ does not express common belief in rationality, we conclude that b_1 must entail that player 1 believes that player

2 does not deem the belief hierarchy inducing $c_1^\infty(timid)$ possible (that is, that belief hierarchy cannot be in the support of b_j). But only choice *timid* is rational for every belief hierarchy where 1 believes that 2 does not deem $c_1^\infty(timid)$ possible ($c_1^\infty(timid)$ would need to have at least probability $\frac{1}{2}$ so that choosing *bold* could be rational for player 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 assigns probability 0 to the belief hierarchy inducing $c_1^\infty(timid)$. 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 induce $c_1^\infty(timid)$. This, however, is a contradiction since we have seen that the belief hierarchy inducing $c_1^\infty(timid)$ 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.

Bibliography

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

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

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

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

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

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

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

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, 2001: Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*, **116 (1)**, 55–79.

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

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

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

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

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

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

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

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

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.

Huber, P. J., 1981: *Robust Statistics*. New York: John Wiley & Sons.

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

Jagau, S., and A. Perea, 2018: Common belief in rationality in psychological games, under review, submitted version at: https://drive.google.com/file/d/1AfaCaXqx5zC95nRBXx7KTpdic0f1JVD9/view.

Kantorovich, L. V., and G. S. Rubinstein, 1958: On a space of completely additive functions. *Vestnik Leningrad. Univ*, **13 (2)**, 52–59.

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.

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

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

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

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.