Reasoning in psychological games: When is iterated elimination of choices enough?*

Niels J. Mourmans†
Maastricht University

EPICENTER Working Paper No. 20 (2019)

Abstract

The framework of psychological game theory has allowed for the modelling of a wide range of belief-dependent motivations. At the same time, analysing psychological games can get complex rather quickly due to the fact that higher-order beliefs may enter the utility functions. As a result, some nice properties of traditional games fail to carry over to psychological games in general. This includes the failure of the iterated elimination of strictly dominated choices (IESDC) to always exactly characterize the choices that are rationally played under belief hierarchies expressing common belief in rationality. In this paper we characterize the families of two-player expectation-based psychological games for which IESDC yields exactly the choices that are rationally played under common belief in rationality. We characterize these games based on which orders of beliefs are directly utility-relevant for a decision-maker. In total we identify three cases. Two of these are relatively trivial: (*i*) the decision-maker's utility depends on a single, even order of belief and (*ii*) the decision-maker's utility and her opponent's utility depend on a single order of belief. We also identify a third, non-trivial case. Our novel notion of causality diagrams, which capture those orders of beliefs that are (indirectly) utility-relevant, is used to obtain our results.

JEL Classification: C72, D03, D83

Keywords: Psychological games, Epistemic game theory, Common belief in rationality, Iterative elimination procedure

1 Introduction

From traditional game theory we have become familiar with reasoning about interactive scenarios where individuals care about material payoffs. However, in many real-life scenarios individuals do not only have preferences that are rooted exclusively in the outcomes of the game. Rather, they are also often motivated by the beliefs and intentions of themselves and others. These types of belief-dependent motivations cannot be captured by the traditional framework of game theory. As a response, the more general framework of *psychological game theory* was introduced by Geanakoplos et al. (1989) and further developed and refined by Battigalli and Dufwenberg (2009). This framework has allowed for the modelling (and experimental testing) of a wide range

*I wish to thank my supervisors Andrés Perea and Elias Tsakas for their many useful comments and support throughout this research project.

†Department of Quantitative Economics, School of Business and Economics, Maastricht University, 6200 MD Maastricht, THE NETHERLANDS; EPICENTER, School of Business and Economics, Maastricht University, 6200 MD Maastricht, THE NETHERLANDS. Email: n.mourmans@maastrichtuniversity.nl

of belief-dependent motivations, including: intention-based reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010), frustration and anger (Battigalli et al., 2015), surprise (Khalmetski et al., 2015), deception and lying behaviour (Battigalli et al., 2013; Dufwenberg and Dufwenberg, 2018; Gneezy et al., 2018), guilt (Dufwenberg, 2002; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007; Attanasi et al., 2013; Attanasi et al., 2016) and social norm conformity (Li, 2008; Charness et al., 2019). While psychological games can introduce very interesting phenomena, they can also be noticeably hard to analyse, certainly compared to traditional games.

A well-established notion that is used to predict behaviour in traditional game theory is the basic concept of common belief in rationality (Spohn, 1982; Brandenburger and Dekel, 1987; Tan and Werlang, 1988). In traditional games, common belief in rationality is appealing for two reasons. Conceptually, it allows for a one-person perspective on a game, as opposed to Nash equilibrium. This means that when making a choice, a player forms beliefs in her mind about what her opponent will choose. She also forms beliefs about what her opponent believes she will choose. And so on. Based on such individual reasoning, the player reaches her decision. Practically, common belief in rationality is also an intuitive notion to use and straightforward to compute in traditional games due to its characterization in terms of iterated elimination of strictly dominated choices (IESDC). It thus becomes a natural and important question to ask to what extent the IESDC-procedure is able to characterize rational choices under common belief in rationality.

In this paper, we will explicitly focus on this question. We will consider the question for a particular class of psychological games. As argued in Jagau and Perea (2018), most applications of psychological game theory are *expectation-based psychological games*. In such games, players in a game care only about higher-order expectations. These are sequences of probability distributions that summarize some, but not all, aspects of a belief hierarchy. One nice property that carries over from traditional games to such expectation-based psychological games is the finite matrix representation of a psychological game. Because of this, such games behave very much like traditional games. Moreover, a finite matrix representation is essential in defining procedures such as IESDC.[1]

Despite the resemblance to traditional games, there are examples of expectation-based psychological games where the IESDC-procedure *fails* to characterize rational choices under common belief in rationality (see for instance Jagau and Perea (2018)). Here we will shed light on the matter of why the IESDC-procedure may fail to characterize common belief in rationality in certain scenarios and why it actually does give an exact characterization in others. We do so by exactly identifying those families of expectation-based psychological games where the IESDC-procedure does give a characterization of rational choices under common belief in rationality. By doing so, we not only identify those families of psychological games that are on a similar level of complexity in terms of reasoning as traditional games. We also point out what can make the other families of psychological games so difficult to reason about, both from the point of view of the player as well as that of the analyst. Our analysis in this paper focuses on two-player expectation-based psychological games in a static environment without updating of beliefs.

In Theorem 1 we show that all rational choices under common belief in rationality must necessarily survive the IESDC-procedure. The other direction however does not need to hold.

To briefly illustrate how the IESDC-procedure can fail to characterize rational choices under common belief in rationality, consider the introductory example of an expectation-based psycho-

[1]There are exceptions in psychological game theory that are not expectation-based psychological games. These include modelling preferences regarding anxiety (Caplin and Leahy, 2004) and suspense (Caplin and Leahy, 2001; Ely et al., 2015). To model such preferences, we need more information than just the higher-order expectations.

Table 1: *Introductory example*

Player 1's extreme second-order expectations

	(c,a)	(c,b)	(d,a)	(d,b)
a	0	0	0	0
b	1	0	1	1

Player 1's utilities

Player 2's extreme first-order expectations

	a	b
c	0	0
d	0	1

Player 2's utilities

logical game in Table 1. Here we have two players: player 1 and player 2. Player 1 has alternatives a and b to choose from, whereas player 2 can choose between options c and d. Player 2's decision problem is as in a traditional game: she cares only about what player 1 will do. This is represented by the lower matrix. Player 1's utility however depends on her full second-order expectation. That is, her expectation about what player 2 is going to do is relevant for her decision, which is her first-order expectation. Additionally however, she cares about what player 2 expects player 1 (herself) to do. These two expectations, one of which is a higher-order expectation, form player 1's second-order expectation. If player 1 chooses a, she always receives a utility of 0. If on the other hand she chooses b, she receives a utility of 0 in case she expects player 2 to choose c while expecting player 2 to believe player 1 will choose b. In all remaining extreme cases of second-order expectations player 1 receives a utility of 1 when choosing b. Player 1's decision problem is depicted by the upper matrix. It is clear here that no choice for player 1 or player 2 is strictly dominated in the relevant decision problem. The IESDC-procedure would therefore not eliminate any choice for any player. However, choice a for player 1 can never be optimal under a belief hierarchy expressing common belief in rationality. Choice a is only optimal under the extreme second-order expectation (c,b), but choice c is never optimal for player 2 given that she expects player 1 to choose b.

The game in Table 1 is part of a particular family of games. Namely one in which one player's utility directly depends on her first-order beliefs and the utility of the other depends on her first-order and second-order beliefs. We identify the different families of expectation-based psychological games based on the orders of beliefs that are directly relevant in shaping the belief-dependent motivations of a decision-maker. We call these utility-relevant orders of beliefs or orders of belief in which the utility is variable. For instance, when modelling simple guilt, whatever a player believes about her opponent's choice, her first-order belief, is irrelevant. However, what the player believes about her opponent's first-order beliefs, which is part of her second-order belief, is important. The utility-relevant order of belief for modelling guilt would be the second order of belief. As another example, the game in Table 1 then belongs to the family of games where player 1's utility depends on her first and second orders of belief and player 2's utility is variable only in her first order of belief.

In this paper we characterize those families of expectation-based psychological games where the IESDC-procedure always characterizes exactly the choices that can rationally be made under common belief in rationality. Take the perspective of player 1. The main theorem (Theorem 2) establishes that the IESDC-procedure *always* characterizes rational choices under common belief in rationality for player 1 if and only if at least one of the following three conditions holds: (i) the utility of player 1 is variable in a single, even order; (ii) the utility of player 1 is variable in a single

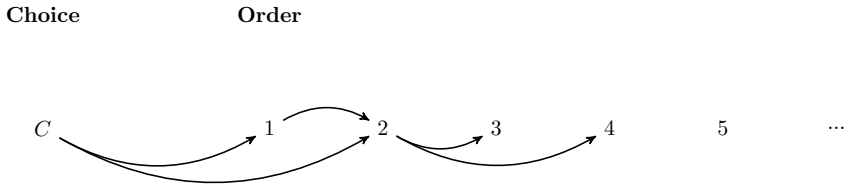


Figure 1: *Causality diagram of player* 1 *in Table 1*

order of belief and the utility of player 2 is variable in a single order of belief as well; (*iii*) player 1's utility is only variable in odd orders and player 2's utility is variable in a single, even order of belief z, such that there is no pair x, y of utility-relevant orders for player 1 and no integer n with $x + n \cdot z = y$. The game in Table 1 does not belong to any family of games described here. An important observation can be made from this result. That is, if players care about material payoffs, cases $(i) - (iii)$ boil down to traditional games where expected utility only depends on first-order beliefs. In all other cases that involve material payoffs one has to go beyond the IESDC-procedure to exactly characterize rational choices under common belief in rationality.

In order for a particular choice to be rational, restrictions then need to apply to the orders of beliefs that are utility-relevant. Under strategic reasoning, it makes sense to assume that the players to which these utility-relevant orders pertain play rationally as well. Similarly, these players as well may have belief-dependent motivations, which are rooted in *their* higher-order beliefs. Then, in order for the decision-maker to believe in the players' rationality at her utility-relevant orders, further restrictions need to be imposed on even higher orders of beliefs. And so on. In the end, we obtain a sequence of orders of beliefs that satisfy all aforementioned restrictions. For player 1 in the introductory example of Table 1 we can illustrate this via a diagram, as depicted in Figure 1. Player 1's utility is variable in her first-order and second-order expectations. These can be directly derived from her first-order and second-order beliefs respectively. For a particular choice C of player 1 to be optimal, restrictions thus need to be imposed on the first-order and second-order beliefs. This is why we have arrows from C to orders 1 and 2 in Figure 1. Player 2's utility is variable only in her first-order expectation. In order for player 1 to believe in player 2's rationality at her already restricted first-order belief, further restrictions need to be imposed on the second-order belief. This is why we have an arrow from order 1, which refers to a belief about player 2's choices, to order 2 in Figure 1. Order 2 refers to a belief about player 1's choices again. For player 1 not to question her own rationality at the second order of belief given the restrictions that have already been imposed on that order of belief, further restrictions are required on the third and fourth orders of belief. Hence the arrows from order 2 to orders 3 and 4. We can continue establishing such arrows indefinitely. Connected arrows together constitute a path in this diagram.

We refer to a diagram like in Figure 1 as a decision-maker's *causality diagram*. Under common belief in rationality, the causality diagram then captures those steps of reasoning of a decision-maker that are directly or indirectly relevant for rationalizing a particular choice. As is for instance the case with order 2 in Figure 1, in a causality diagram the same order of belief may be reached by multiple paths. In the diagram there, order 2 is reached via the path $(0, 2, ..)$, but also via the path $(0, 1, 2, ...)$. If the same order of belief is reached by multiple paths, it implies that in order to rationalize a particular choice under common belief in rationality, not questioning rationality at different orders of beliefs will require different restrictions on the same higher-order belief. If these restrictions are contradictory, then the choice in question cannot be rational under a belief hierarchy expressing common belief in rationality. Exactly this friction is what the IESDC-procedure cannot

pick up on. As a result, the IESDC-procedure may allow for choices that are not rational under common belief in rationality.

If the same order of belief is reached by multiple paths in a causality diagram of a player, we say that two paths in the diagram overlap. If a causality diagram is completely free of any overlapping paths, then also no contradictory restrictions on the same order of belief can occur. The main theorem of this paper describes exactly these cases by means of families of expectation-based psychological games. Two of the three cases are trivial in the sense that the causality diagram only has a single path. Moreover, none of the cases include scenarios where both players care about materialized outcomes and at least one player has some belief-dependent motivation.

The remainder of this paper is structured as follows. Section 2 presents a definition of static psychological games and common belief in rationality in such games. Section 3 discusses the concept of higher-order expectations, expectation-based psychological games and families of expectation-based psychological games based on utility-variant orders. Section 4 discusses the IESDC-procedure and its problems in psychological games. Moreover, we state here the main result of the paper: Theorem 2. In Section 5 we introduce the notion of causality diagram to visualize reasoning in expectation-based psychological games. Section 6 is fully dedicated to the proof of Theorem 2. Some parts of the proof are moved to the Appendix that accompanies this paper. However, Section 6 illustrates these parts of the proof by means of examples. Finally, we end this paper with some concluding remarks in Section 7.

2 Preliminaries

In this section we discuss the framework of psychological game theory in general. Moreover, we will define the reasoning concept of common belief in rationality in this framework, which will play a central role in this paper.

2.1 Psychological games

In a traditional setting, a player's experienced, ex-post utility depends only on her opponents' choices. The player's utility of making a particular decision then depends only on her first-order belief of what she expects her opponents to choose. As opposed to this, a player's expected utility in a *psychological* game can explicitly, and non-linearly, depend on any order of belief or even the entire *belief hierarchy*. In order to formally discuss the framework of a psychological game, we should therefore clarify what a belief hierarchy formally is and what types of belief hierarchies we will be restricting to in this paper. A belief hierarchy b_i for a player i represents an infinite chain of beliefs. The first element in this chain represents the first-order belief about the opponents' choices, the second element represents the second-order belief about the opponents' choices combined with the opponents' beliefs about their opponents' choices and the third represents the third-order belief about the combination of opponents' choices, opponents' first-order beliefs and the opponents' second-order beliefs. And so on.

Following Brandenburger and Dekel (1993), we formally define beliefs in spaces of uncertainty. Consider any Polish space S. Let $\Delta(S)$ be the set of probability measures on the Borel σ-field over the space of uncertainty S. Finally, endow $\Delta(S)$ with the topology of weak convergence. Then $\Delta(S)$ is a Polish space as well. We are considering two-player psychological games throughout this paper. The primitive space of uncertainty for player i in a two-player game is the set of the opponent's choices C_j. We can recursively define

$$
\begin{aligned}
X_i^1 &:= C_j \\
X_i^2 &:= X_i^1 \times \Delta(X_j^1) \\
&\vdots \\
X_i^n &:= X_i^{n-1} \times \Delta(X_j^{n-1}) \\
&\vdots
\end{aligned}
$$

Then the set of all possible belief hierarchies for player i is $\tilde{B}_i := \times_{n=1}^{\infty} \Delta(X_i^n)$. Each belief hierarchy b_i is a vector of (higher-order) beliefs $(b_i^1, b_i^2, ...)$, where the $n-th$ order belief of player i is a probability distribution $b_i^n \in \Delta(X_i^n)$. In the current set-up a belief hierarchy $b_i \in \tilde{B}_i$ may be incoherent in the sense that an n-th order belief may contradict what is stated by the $(n-1)$-th order belief. When defining psychological games, we assume a player's belief hierarchy cannot show such incoherences. More formally, we define *coherency* as follows.

Definition 1. *A belief hierarchy $b_i = (b_i^1, b_i^2, ...)$ expresses **coherency** if for every $n > 1$ we have*

$$marg_{X_i^{n-1}} b_i^n = b_i^{n-1}.$$

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

From Brandenburger and Dekel (1993)'s Proposition 1 we know there exists a homeomorphism $f_i : \tilde{B}_i(1) \to \Delta(C_j \times \tilde{B}_j)$. Thus a coherent belief hierarchy can be identified with a probability distribution over the possible combinations of the opponent's choices and belief hierarchies. Next to expressing coherency, a player can also believe her opponent expresses coherency, believe that her opponent believes *her* opponent expresses coherency, and so on. This restricts the set of belief hierarchies we will consider further. We can recursively define such sets of belief hierarchies:

$$\tilde{B}_i(k) = \{b_i \in \tilde{B}_i(k-1) | f_i(b_i)(C_j \times \tilde{B}_j(k-1)) = 1\},\ k \geqslant 2.$$

Consider the set $B_i = \bigcap_{k \geqslant 0} \tilde{B}_i(k)$. We say a belief hierarchy b_i expresses *coherency and common belief in coherency* if $b_i \in B_i$. Throughout this paper, whenever we refer to a belief hierarchy b_i, we assume it to be a belief hierarchy in B_i, even if it is not directly stated as such. Moreover, every $b_i \in B_i$ can be identified with a probability distribution over $C_j \times B_j$ through the homeomorphism f_i as was used in defining $\tilde{B}_i(k)$ for any $k \geqslant 1$. We will make use of this fact multiple times throughout the paper.

With these elements in place, we can now give a formal definition of a psychological game. We follow here the approach taken by Jagau and Perea (2017).

Definition 2. *A **psychological game** is a tuple $G = (C_i, B_i, u_i)_{i \in I}$, where C_i is the finite set of choices for player i[2], B_i denotes the set of belief hierarchies expressing coherency and common belief in coherency and*

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

is player i's (measurable) utility function.

[2] C_i may well be a singleton set, indicating a situation where player i does not have any choices to make but where his beliefs matter for the utilities of other players.

By this definition, we capture the idea that player i's utility depends explicitly on her full belief hierarchy. Formally speaking a psychological game is a generalisation of a traditional game, since the utility function in a traditional game exclusively depends on first-order beliefs. Moreover, utilities in a traditional game always depend linearly on first-order beliefs. This is not true for psychological games in general, where utilities can depend non-linearly on the full belief hierarchy. Definition 2 differs from definitions used in the seminal work by Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009). Under these two definitions, utility also still explicitly depends on the opponent's choices. In case of the latter approach, utility moreover explicitly depends on the opponent's belief hierarchies as well. These two elements are helpful in visually distinguishing between preferences over outcomes and belief-dependent motivations. However, as Jagau and Perea (2017) point out, all these approaches are essentially equivalent. This can be seen by noting that a belief hierarchy can be identified by a probability distribution over the combinations of the opponent's choices and *her* belief hierarchies $C_j \times B_j$. Hence, a belief hierarchy already includes a conjecture about the opponent's choices and opponent's belief hierarchies. In terms of utility that is deemed relevant at the moment of making a decision, as opposed to ex-post experienced utility, all approaches are thus equivalent in an expected utility framework.

2.2 Common belief in rationality in psychological games

The central theme of this paper revolves around elimination procedures that characterize common belief in rationality in psychological games. Much like in traditional games, common belief in rationality conveys the idea that nowhere in one's belief hierarchy the rationality of any player is questioned. The notion of common belief in rationality in psychological games was first presented in Battigalli and Dufwenberg (2009) by their discussion of common strong belief in rationality in dynamic psychological games (the equivalent of this concept in static games is common belief in rationality). Later it was the focal point in Jagau and Perea (2017). Our definition of the concept follows that of the latter paper.

Common belief in rationality in a psychological game can be defined recursively. First we consider an *optimal* choice given any belief.

Definition 3. *Consider a psychological game G. A choice c_i is* ***optimal*** *for a belief hierarchy $b_i \in B_i$ if for all $c_i' \in C_i : u_i(c_i, b_i) \geqslant u_i(c_i', b_i)$.*

Let $RB_i := \{(c_i, b_i) \in C_i \times B_i | c_i \text{ optimal given } b_i\}$ be the set of combinations of choices and belief hierarchies where the choice is optimal for the belief hierarchy. Then we can define what it means to believe in an opponent's rationality. To this end, recall that every belief hierarchy $b_i \in B_i$ is homeomorphic to a probability distribution in $\Delta(C_j \times B_j)$.

Definition 4. *Consider a belief hierarchy $b_i \in B_i$ for some player i in G. Belief hierarchy b_i is said to express* ***belief in the opponent's rationality*** *if $b_i(RB_j) = 1$.*

In line with Spohn (1982), Bernheim (1984), Pearce (1984) and Tan and Werlang (1988) for standard games, we can iterate this definition to get the notion of common belief in rationality in a psychological game (Jagau and Perea, 2017).

Definition 5. *Consider $b_i \in B_i$ for some player i. Define $B_i(1) = \{b_i \in B_i | b_i(RB_j) = 1\}$. If $b_i \in B_i(1)$, we say b_i expresses* ***1-fold belief in rationality****.*
For every $k \geqslant 1$, define $B_i(k) = \{b_i \in B_i(k-1) | b_i \in \Delta(C_j \times B_j(k-1))\}$. We say b_i expresses up to ***k-fold belief in rationality*** *if $b_i \in B_i(k)$.*
If for every $k \geqslant 1$, b_i expresses up to k-fold belief in rationality, we say b_i expresses ***common belief in rationality****.*

We assume here that the events of expressing up to k-fold belief in rationality are measurable. Finally, let a rational choice under k-fold belief in rationality and common belief in rationality respectively be denoted as follows.

Definition 6. *Consider a choice $c_i \in C_i$. We say c_i is **rational under up to k-fold belief in rationality** if it is optimal for some belief hierarchy $b_i \in B_i$ that expresses up to $k-fold$ belief in rationality. We say c_i is **rational under common belief in rationality** if it is optimal for some belief hierarchy $b_i \in B_i$ that expresses common belief in rationality.*

This definition of common belief in rationality and rational choice under a belief hierarchy that expresses common belief in rationality is applicable to any class of psychological games.

3 Higher-order expectations

In this section we will be looking at a particular subclass of psychological games: expectation-based psychological games. Furthermore, we will discuss how we will distinguish between different families of psychological games. These families will be defined by the orders of belief that are directly relevant for players' utilities in a given expectation-based psychological game.

3.1 Expectation-based psychological games

When modelling emotions or other-regarding preferences, we typically do not use all information contained in higher-order beliefs in the utility functions (Jagau and Perea, 2018). Instead, we use *expectations* about expectations, which can be derived from (higher-order) beliefs. For instance, in modelling surprise-related preferences we only care about a player's *expectation* about her co-players' first-order beliefs in determining her psychological payoff. To illustrate this, consider the following example.

Example 1 (Surprising student by means of exam)**.** *You are Ann's high school teacher in economics. You noticed that Ann's focus during classes has been lacking. Therefore, you wish to give her a wake-up call by surprising her (and the rest of the small class in the process). To this end, at the end of a given school week you vaguely announce to the class that next Monday an exam might be given. You can surprise Ann in two ways. You either give the exam on Monday while Ann does not expect one or you do not give the exam on Monday while she did expect you to give one. Surprising Ann gives you a feeling of psychological satisfaction. Either form of surprise is equally satisfactory to you. Even though you wish to surprise Ann, you also do not want her to fail.*

How do we model your utility as a teacher in the above example? Let us have $I = \{y, a\}$, with $y = you$ and $a = Ann$. For the choice-problem presented to us before Monday we moreover have $C_y = \{exam, no\ exam\}$ for your choice-set as the teacher and $C_a = \{study, not\ study\}$ as the choice set for Ann.

The utility you receive from your decision before Monday depends on two factors: the probability with which you believe Ann will study for the possible exam and the probability with which you expect Ann to believe you will actually give the exam. Let us say that any form of surprise gives you as a teacher one extra unit of utility, Ann failing the exam makes you lose a unit of utility and Ann succeeding makes you gain a unit of utility. If we assume your utility to be additive in these two different components, we can describe your expected utility of giving an exam on Monday by the following relation

$$u_y(exam, b_y) = (1 - \int_{C_a \times B_a} b_a^1(exam) db_y) + (2 \cdot b_y^1(study) - 1).$$

Table 2: *Surprise exam game*

	Extreme Second-order expectations			
	$(study, exam)$	$(study, no\ exam)$	$(not\ study, exam)$	$(not\ study, no\ exam)$
$Exam$	1	2	-1	0
$No\ exam$	1	0	1	0

The expected utility function above depends on a summary statistic of the second-order belief induced by your belief hierarchy b_y. The second component in the expected utility function corresponds to your first-order belief that Ann will study. The integral measure on the other hand represents your *expectation* of Ann's first-order belief about your choice, induced by your belief hierarchy b_y. This summary statistic is called your *second-order expectation* in this setting. For instance, we can define Ann's *first-order expectation* that you will give an exam as $e^1_a[b_a](exam) := b^1_a(exam)$. Then your second-order expectation that Ann will study and that she believes you will give an exam is given by

$$e^2_y[b_y](study, exam) := \int_{\{study\}\times B_a} e^1_a[b_a](exam)db_y = \int_{\{study\}\times B_a}\int_{\{exam\}\times B_y} db_a db_y.$$

Notice that the above expectation is a joint probability measure. In fact, every second-order expectation for you as a teacher is a joint probability measure $e^2_y \in \Delta(C_a \times C_y)$. We can directly represent your utility of choosing to give an exam as a function of your second-order expectation induced by your belief hierarchy b_y: $u_y(exam, e^2_y[b_y])$. Similarly, the expected utility of not giving an exam can be represented by $u_y(no\ exam, e^2_y[b_y]) = \int_{C_a\times B_a} e^1_a[b_a](exam)db_y$ (if you do not give an exam, you will not see Ann succeed or fail). Since both C_a and C_y are finite sets of choices, we have that $C_a \times C_y$ is finite as well. The set of second-order expectations then has finitely many extreme points. Consequently, the resulting utility for you as a teacher for all possible extreme second-order expectations can be represented in finite-matrix form as in Table 2.

Following Jagau and Perea (2018), we can define any *higher-order expectation* recursively. To define any particular class of psychological games where utilities depend on higher-order expectations, this is a useful tool.

In a general psychological game G, let us start with the *first-order expectation* $e^1_i[b_i]$ for a player i. This is simply the first-order belief of player i:

$$e^1_i[b_i] := b^1_i \in \Delta(C_j).$$

The *second-order expectation* of player i that player j will choose c_j while believing that player i will choose c_i is subsequently defined as follows:

$$e^2_i[b_i](c_j, c_i) := \int_{\{c_j\}\times B_j} e^1_j[b_j](c_i)db_i = \int_{\{c_j\}\times B_j}\int_{\{c_i\}\times B_i} db_j db_i.$$

As noted in our surprise exam example (Example 1), a second-order expectation is a joint probability measure $e^2_i[b_i] \in \Delta(C_j \times C_i)$. We can recursively define any higher-order expectation following this construction. Note however that second-order expectations are defined over the Cartesian product of two choice sets. As the orders of higher-order expectations increase, this Cartesian product will contain more and more elements as well. For the sake of clarity in our notation, we

will therefore define the following sets recursively:

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

Each $w_i^k \in W_i^k$ thus has k components and represents a combination of your opponent's and your own choices. Throughout this paper, we will utilise the following identification of $w_i^k \in W_i^k$ at times as well: $w_i^k = (c_j, w_j^{k-1})$. A k-th order expectation is then defined as a probability measure over W_i^k.

Definition 7 (Jagau and Perea (2018))**.** *Consider a two-player psychological game G with a player i and a player j and let b_i be the belief hierarchy for player i. Let $e_i^1[b_i] := b_i^1$ be the first-order expectation for player i given b_i. For $k \geqslant 2$, the **k-th order expectation** $e_i^k[b_i] \in \Delta(W_i^k)$ of player i given belief hierarchy b_i is defined as*

$$e_i^k[b_i](w_i^k) := \int_{\{c_j\} \times B_j} e_j^{k-1}[b_j](w_j^{k-1}) db_i, \text{ where } w_i^k = (c_j, w_j^{k-1}).$$

In the integral b_i serves as a probability measure over $C_j \times B_j$, similar to how it was used in Section 2.2.

We can capture all possible k-th order expectations in the set $E_i^k := \Delta(W_i^k)$. This allows us to define utilities that depend explicitly on k-th order expectations by $u_i : C_i \times E_i^k \to \mathbb{R}$. Notice here that $e_i^k[b_i]$ for any $k > 1$ given a belief hierarchy b_i also contains the lower-order expectations induced by said belief hierarchy. That is, we have that $marg_{W_i^{k-1}} e_i^k[b_i] = e_i^{k-1}[b_i]$. Much like beliefs, we thus obtain a *hierarchy of expectations* $e_i[b_i] := (e_i^1[b_i], e_i^2[b_i], ...)$ induced by a belief hierarchy b_i.

Two points are worthwhile to elaborate on here leading to the upcoming Definition 8. First, the mapping from the set of belief hierarchies to the set E_i^k is surjective but non-injective. For every k-th order expectation, there is a belief hierarchy that induces it. However, a given k-th order expectation may be induced by multiple belief hierarchies. This is illustrated in Table 3, where we depict the second-order beliefs of two possible belief hierarchies for you as the teacher. The second-order expectations induced by the two belief hierarchies are equal, whereas the second-order beliefs are not. Indeed, in b_y^2 you are certain about Ann's belief and believe that Ann is uncertain

Table 3: *Illustration of belief hierarchies and second-order expectations for Example 1*

Your second-order beliefs	b_y^2	=	$(study, b_a^1)$
	$b_y^{2\prime}$	=	$\frac{1}{2}(study, b_a^{1\prime}) + \frac{1}{2}(study, \hat{b}_a^1)$
Ann's first-order beliefs	b_a^1	=	$\frac{1}{2}(exam) + \frac{1}{2}$(no exam)
	$b_a^{1\prime}$	=	$exam$
	$\hat{b}_a^1$	=	$no\ exam$
Your second-order expectations	$e_i^2[b_y]$	=	$\frac{1}{2}(study, exam) + \frac{1}{2}(study, no\ exam)$
	$e_i^2[b_y']$	=	$\frac{1}{2}(study, exam) + \frac{1}{2}(study, no\ exam)$

Table 4: *Surprise exam game with a mean teacher*

	Your extreme second-order expectations			
	$(study, exam)$	$(study, no\ exam)$	$(not\ study, exam)$	$(not\ study, no\ exam)$
Exam	0	0	0	1
No exam	1	0	0	0

Teacher's utilities

about your choices. In $b_y^{2\prime}$ you are uncertain about Ann's belief but believe that Ann is certain about your choice.

Second, recall again from Section 2.1 that any belief hierarchy b_i can be represented by a probability distribution in $\Delta(C_j \times B_j)$. Take any two belief hierarchies $b_i, b_i' \in B_i$. The **convex combination** $\lambda b_i + (1-\lambda)b_i'$ for any $\lambda \in [0,1]$ is then the belief hierarchy that puts probability $\lambda f_i(b_i)(E) + (1-\lambda) f_i(b_i')(E)$ to every measurable $E \subseteq C_j \times B_j$. With the previous two points in mind, we are now in a position to formally define a psychological game where expectations instead of beliefs matter explicitly for utilities.

Definition 8 (Jagau and Perea, 2018)**.** *We call a two-player psychological game $G = (C_i, B_i, u_i)_{i \in I}$ an **expectation-based psychological game** if, for both players i and all choices $c_i \in C_i$,*

(i) $e_i[b_i] = e_i[b_i']$ implies $u_i(c_i, b_i) = u_i(c_i, b_i')$

(ii) utility is linear in the beliefs hierarchies: $u_i(c_i, \lambda b_i + (1-\lambda)b_i') = \lambda u_i(c_i, b_i) + (1-\lambda) u_i(c_i, b_i')$, for all $\lambda \in [0,1]$.

The second condition is that of **belief linearity**. This condition states that the expected utility given c_i and the convex combination of two belief hierarchies b_i and b_i' has to be equal to the convex combination of the two expected utilities induced by the choice $c_i \in C_i$ and by the belief hierarchies b_i and b_i'. Finally, we can impose a last, natural condition.

Definition 9. *A psychological game $G = (C_i, B_i, u_i)_{i \in I}$ is **belief-finite** if there is some $n \geqslant 1$ such that for every choice $c_i \in C_i$, and every two belief hierarchies $b_i, \hat{b}_i \in B_i$ with $b_i^n = \hat{b}_i^n$ we have that $u_i(c_i, b_i) = u_i(c_i, \hat{b}_i)$.*

In words, belief-finiteness means that utility depends only on finite orders of beliefs. Belief-finiteness allows us to have a finite representation of an expectation-based psychological game in matrix form. This is because there are finitely many extreme higher-order expectations under belief-finiteness for a player i to consider. These extreme higher-order expectations are represented in the columns of the matrix (see for instance Table 2). The choices, as traditionally is the case, are found in the rows. For the remainder of the paper, we assume every expectation-based psychological game we will be dealing with is belief-finite.

Finally, note that the example in Table 2 assumes that your utility as a teacher by giving the exam is additively separable in wanting to surprise Ann and wanting her to pass the exam. However, by definition of an expectation-based, belief-finite psychological game, utility does not always have to be additive in the different higher-order expectations. To this end, reconsider Example 1, but now assume you are a mean teacher. That is, you wish to surprise Ann by giving the exam if she does not study, and you wish to surprise Ann by not giving an exam if she does study. Any other scenario does not interest you. This non-additively separable psychological game is illustrated in Table 4.

3.2 Order-variable families of psychological games

The problem that a player in any belief-finite, expectation-based psychological game faces can be thought of as a decision problem. Generally, a *decision problem* can be defined by a triple $D = (C, X, v)$. In this triple, C refers to a finite set of choices, X is a finite set of states and $v : C \times X \to \mathbb{R}$ is a Bernoulli utility function. A choice $c \in C$ is then *optimal* in D if there is a belief $b \in \Delta(X)$ such that

$$\sum_{x \in X} b(x) \cdot v(c, x) \geqslant \sum_{x \in X} b(x) \cdot v(c', x), \forall c' \in C.$$

In a belief-finite, expectation-based psychological game where utilities only depend up to order k the set of states X would then refer to W_i^k for $k \geqslant 1$. The utility $v_i(c_i, w_i^k)$ then refers to the utility experienced from choosing c_i while being in state w_i^k.

We can define families of expectation-based psychological games depending on which orders of beliefs are of direct relevance to a player's preferences.

Definition 10. *Let $i \in \{1, 2\}$. Take a belief finite, expectation-based psychological game G, where player i's utility function can be summarized by $v_i : C_i \times W_i^n \to \mathbb{R}$. If $v_i(c_i, w_i^n) \neq v_i(c_i, \hat{w}_i^n)$ for some $c_i \in C_i$ and some w_i^n and $\hat{w}_i^n$ that only differ in the m-th order, we say v_i is* ***variable in the m-th order****.*
By $\mathcal{G}(N_1, N_2)$ we denote the ***family of psychological games*** *in which player 1's and player 2's utility-variable orders are specified by N_1 and N_2 respectively, with $N_1, N_2 \subseteq \mathbb{N}$.*

When we refer to '(directly) utility-relevant' orders of belief for a player i in the remainder of the paper, we always mean the orders of belief in which player i's utility is variable.

4 Iterated elimination of strictly dominated choices

In this section we will discuss the procedure of iterated elimination of strictly dominated choices in psychological games. Unlike traditional games, this procedure does not always characterize the rational choices a player can make under common belief in rationality. However, there are some families of games for which this relationship does hold. This leads us to state the main result of this paper, captured in Theorem 2. In the second part of this section we provide intuition on why the elimination procedure may fail in its characterization of rational choices under common belief in rationality.

4.1 Iterated elimination of strictly dominated choices in psychological games

In traditional games, choices that are rational under some belief hierarchy expressing common belief in rationality are characterized by iteratively eliminating strictly dominated choices. We say a choice $c \in C$ is *strictly dominated* in a decision problem $D = (C, X, v)$ if there is a randomized choice $r \in \Delta(C)$ such that

$$v(c, x) < \sum_{c' \in C} r(c') \cdot v(c', x), \forall x \in X.$$

Under the set-up presented above, iterative elimination of strictly dominated choices then means that each round of eliminating choices induces a new decision problem for a decision-maker. In each round, those choices are eliminated that are never optimal in the given decision problem. We can use a result by Pearce (1984) for this.

Lemma 1 (Pearce's Lemma)**.** *Consider a decision problem $D = (C, X, v)$. Then, $c \in C$ is optimal in D if and only if c is not strictly dominated in D.*

Pearce's original lemma defines the set of states X as the set of choices C_j of the opponent j of a player in a traditional, two-player game. But his proof technique can be used to prove Lemma 1 as well.

In a belief-finite, expectation-based psychological game where utilities only depend up to order n the set of states X would refer to W_i^n for $n \geqslant 1$. The utility $v_i(c_i, w_i^n)$ refers to the utility experienced from choosing c_i while being in state w_i^n. The procedure of iterative elimination of strictly dominated choices (IESDC) is then defined as follows.

Procedure 1 (Iterated elimination of strictly dominated choices (IESDC))**.**
Consider a two-player psychological game $G = (C_i, B_i, u_i)_{i \in I}$ which is expectation-based and belief-finite, and in which utilities depend up to the n-th order expectation. For every player i, consider the full decision problem $(C_i^0, W_i^{n,0}, v_i)$, where $C_i^0 := C_i$, $W_i^{n,0} := W_i^n$ and $v_i : C_i \times W_i^n \to \mathbb{R}$ summarizes the utility function u_i.

Step 1
For each player i, define: $C_i^1 = \{c_i \in C_i |\ c_i$ is not strictly dominated in $(C_i^0, W_i^{n,0}, v_i)\ \}$.

For each player i, define: $W_i^{n,1} = \begin{cases} C_j^1 \times C_i^1 \times ... \times C_j^1 & \text{if } n \text{ is odd.} \\ C_j^1 \times C_i^1 \times ... \times C_i^1 & \text{if } n \text{ is even.} \end{cases}$

***Step* $\mathbf{k \geqslant 2}$**
For each players i, define: $C_i^k = \{c_i \in C_i^{k-1} |\ c_i$ is not strictly dominated in $(C_i^{k-1}, W_i^{n,k-1}, v_i)\ \}$.

For each player i, define: $W_i^{n,k} = \begin{cases} C_j^k \times C_i^k \times ... \times C_j^k & \text{if } n \text{ is odd.} \\ C_j^k \times C_i^k \times ... \times C_i^k & \text{if } n \text{ is even.} \end{cases}$

For each players i, define $C_i^\infty = \bigcap_{k \geqslant 1} C_i^k$.

Some explanation is due here. In this procedure we assume that player i only cares for higher-order expectations up to order n (this may or may not include order n). In Step 1, both players i eliminate those choices that are strictly dominated in their respective decision problems. Subsequently we define the resulting sets of combination of choices $W_i^{n,1}$ for each i which are constructed from those sets of choices that are not strictly dominated in the original decision problems. Then, using C_i^1 and $W_i^{n,1}$ we can construct a *reduced decision problem* $(C_i^1, W_i^{n,1}, v_i)$, where $v_i : C_i^1 \times W_i^{n,1} \to \mathbb{R}$. Note that for the identification of the utility function v_i we technically abuse notation in this elimination step. Formally we have $v_i : C_i \times W_i^n \to \mathbb{R}$, but we identify it with a restriction on $C_i^1 \times W_i^{n,1}$. After constructing the reduced decision problem, we repeat the process. The procedure ends when no choices can be eliminated any longer for any of the two players.

This procedure leads us to consider the following theorem for this section.

Theorem 1 (Rational choice under common belief in rationality requires surviving the procedure)**.**
Consider any belief-finite, expectation-based psychological game G, with two players where utilities depend on up to n-th order expectations. Then every choice that is rational under common belief in rationality must necessarily survive IESDC.

Proof. Define by R_i^∞ the set of rational choices player i can make under common belief in rationality. We will prove that $R_i^\infty \subseteq C_i^\infty$. We will do this by showing that $R_i^\infty \subseteq C_i^k$ for every $k \geqslant 1$. This will be done by induction on k. We will start with the case of $k = 1$.

Induction start Take an arbitrary choice $c_i \in R_i^\infty$. This means that c_i is optimal for some belief hierarchy $b_i \in B_i$ that expresses common belief in rationality. By Pearce's Lemma this implies $c_i \in C_i^1$ as well, as c_i being optimal for some belief hierarchy $b_i \in B_i$ among all choices in C_i is exactly the same as c_i not being strictly dominated in the decision problem $(C_i^0, W_i^{n,0}, v_i)$.

Induction step Assume that $R_i^\infty \subseteq C_i^{k-1}$ for some $k \geqslant 2$ for each player i. Now take some $c_i \in R_i^\infty$. Then we have that c_i is optimal for some belief hierarchy $b_i \in \Delta(C_j \times B_j)$ that expresses common belief in rationality, among all choices in C_i. Common belief in rationality implies that player i believes that her opponent player j makes an optimal choice according to a belief hierarchy that expresses common belief in rationality. Hence each $(c_j, b_j) \in \mathit{supp}(b_i)$ is such that c_j is optimal for b_j which expresses common belief in rationality. Thus, $c_j \in R_j^\infty$ by definition and by our induction assumption therefore $c_j \in C_j^{k-1}$.

However, if $b_j \in \Delta(C_i \times B_i)$ expresses common belief in rationality, this implies that player i believes that player j believes player i expresses common belief in rationality and makes an optimal choice accordingly. Hence, each $(c_i', b_i') \in \mathit{supp}(b_j)$ is such that c_i' is optimal for b_i' which expresses common belief in rationality. Therefore, $c_i' \in R_i^\infty$ and by the induction assumption, $c_i' \in C_i^{k-1}$. As a result, the choice c_i we started with must be optimal for a belief hierarchy $b_i \in \Delta(C_j^{k-1} \times \Delta(C_i^{k-1} \times B_i))$.

If we continue this line of reasoning that each player i believes rationality is commonly believed, we get that the $c_i \in R_i^\infty$ we started with in this induction step must be optimal for a belief hierarchy $b_i \in \Delta(C_j^{k-1} \times \Delta(C_i^{k-1} \times \Delta(C_j^{k-1} \times ...)))$. If we take the n-th order expectation induced by this belief hierarchy, we get $e_i^n[b_i] \in \Delta(\underbrace{C_j^{k-1} \times C_i^{k-1} \times ... \times C_j^{k-1}}_{\text{n times}})$ if n is odd and $e_i^n[b_i] \in \Delta(\underbrace{C_j^{k-1} \times C_i^{k-1} \times ... \times C_i^{k-1}}_{\text{n times}})$ if n is even. Thus c_i is optimal for some n-th order expectation $e_i^n[b_i] \in \Delta(W_i^{n,k-1})$ induced by belief hierarchy b_i. By Pearce's Lemma then c_i is not strictly dominated in the decision problem $(C_i, W_i^{n,k-1}, v_i)$. It follows that c_i is then also not strictly dominated in the decision problem $(C_i^{k-1}, W_i^{n,k-1}, v_i)$. Hence, $c_i \in C_i^k$ and since we took an arbitrary $c_i \in R_i^\infty$, we also have that $R_i^\infty \subseteq C_i^k$.

By induction on k, we have that $R_i^\infty \subseteq C_i^k$ for every $k \geqslant 1$, which completes the proof. □

We have thus shown that if a choice is rational under common belief in rationality it must survive the IESDC-procedure. This result holds given a game $G \in \mathcal{G}(N_1, N_2)$ for *any* family of games $\mathcal{G}(N_1, N_2)$, $N_1, N_2 \subseteq \mathbb{N}$. The reverse statement does not always hold. It is not true that those choices that survive the IESDC-procedure are always rational under common belief in rationality (Jagau and Perea, 2017). The example of Table 1 illustrates this. There are however specific families of games where, for *all* games in such a family, the IESDC-procedure does exactly characterize rational choices under common belief in rationality.

Theorem 2. *Consider any family $\mathcal{G}(N_1, N_2)$ of belief-finite, expectation-based psychological games with two players. For every game in $\mathcal{G}(N_1, N_2)$, each choice for player 1 that survives the IESDC-procedure is also a rational choice under common belief in rationality, if and only if, one of the following conditions is true:*

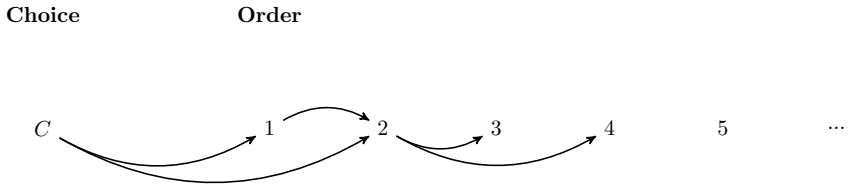


Figure 2: *Causality diagram of player* 1 *in Table 1 (repeated)*

(i) Player 1's utility and player 2's utility are both variable in a single order of belief;

(ii) Player 1's utility is variable in a single even order of belief;

(iii) Player 1's utility is variable only in odd orders of belief and player 2's utility is variable in a single even order of belief z which is such that there is no pair x, y of player 1's utility-variant orders and no $n \in \mathbb{N}$ with $x + n \cdot z = y$.

An important observation we can make here is that if both players care (among others) about material payoffs, the conditions listed in the theorem above reduce to those that specify a traditional game.

In the remainder of this paper, Theorem 2 will take center-stage.

4.2 Illustrating the problem of the IESDC-procedure

Recall the example from the Introduction in Table 1. There is an underlying reason why in this example in Table 1 the IESDC procedure does not give us exactly the choices one can rationally make under common belief in rationality. In short terms, there is an overlap between the orders in which one's utility is variable on the one hand and the orders of beliefs one needs to consider for expressing 1-fold belief in rationality on the other hand. This is also illustrated in the causality diagram for player 1 in Figure 1, repeated here in Figure 2. We will formalize such diagrams in Section 5. In words, we can say the following however. In the diagram in Figure 2, the arrow from player 1's choice C to order "1" indicates that the optimality of player 1's choice depends on her first-order expectation. The same applies for the arrow from her choice to order "2": the optimality of player 1's choice also depends on her second-order expectation. Player 2's utility only depends on her first-order expectation. This is represented by the arrow from order "1" to order "2". And so on. Clearly then, the rationality of player 1's own choice and believing that player 2 makes a rational choice both directly depend on player 1's second-order expectation. Thus for c to be optimal and for player 1 to believe in player 2's rationality, different restrictions need to be imposed on the second-order expectation. In this example these restrictions happen to be in conflict as also explained in the Introduction: in order for choice a to be optimal player 1 in her first-order expectation has to believe that player 2 will choose c while expecting in her second-order expectation that player 2 expects her to choose b. However, in order for choice c to be optimal for player 2, she must expect player 1 to choose a.

Believing in an opponent's rationality is believing in the event that your opponent makes an optimal choice *given* her belief. Belief in the opponent's rationality thus restricts the combinations of choices and belief hierarchies (c_j, b_j) you can consider for the opponent where c_j is optimal specifically for b_j. The first step of the IESDC-procedure however only eliminates choices for your opponent which are never optimal, given *any* belief hierarchy. The second step subsequently only eliminates choices that are never optimal given beliefs that only assign positive probability to choices that are not eliminated in Step 1.

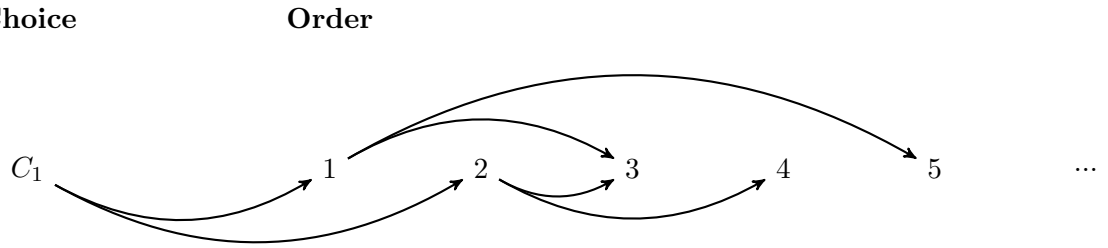


Figure 3: *Causality diagram of player* 1 *in game in* $\mathcal{G}(\{1,2\},\{2,4\})$

This kind of procedure does not sufficiently restrict the second-order expectations one can consider under belief in the opponent's rationality. One may assign positive probability to an extreme second-order expectation whose two components entail choices for an opponent and oneself that did survive Step 1 of the procedure. However, as illustrated via the example in Table 1, we then allow for scenarios in which the player 2's choice in the first component can never be optimal given the conditional probability assigned to player 1's choice that features in the second component. In a traditional game there is no such issue at all, as player 1's utility would only depend on her first-order belief. The rationality of her first-order belief would only depend on her second-order belief. And so on. In this case each step k of eliminating strictly dominated choices corresponds exactly to the reasoning step of expressing up to $(k-1)$-fold belief in rationality. This corresponds also with the observation that traditional games are a special sub-case of case (i).

In the game in Table 1 we have that the utility for player 1 is dependent on her first order and second order of belief, which overlap with the orders of belief that matter for expressing 1-fold belief in rationality. That is, 1-fold belief in rationality is determined by a first-order belief of player 1, rationalized by her second-order belief. In general, this kind of overlap may also occur because of overlap between deeper levels of reasoning. Consider for instance a causality diagram as in Figure 3 for player 1. Here player 1's utility depends on her first-order and second-order expectations, whereas player 2's utility depends on her second-order and fourth-order expectations. Here we see that expressing 1-fold belief in rationality and expressing 2-fold belief in rationality both require restrictions on player 1's third-order expectation, which may be in conflict.

First, there is the event of expressing 1-fold belief in rationality, as the first-order expectation is a conjecture about player 2's choice, which is motivated by player 2's second-order and fourth-order expectations. Hence, a rational first-order belief is explained by player 1's third-order expectation and fifth-order expectation[3]. Second, there is the event of expressing 2-fold belief in rationality, which next to the fourth-order expectation, also depends on the third-order expectation. Namely, the second-order belief is a conjecture about player 1's own choice. The utility of player 1 depends on her first-order and second-order expectation. So in order to rationalize the choices in her second-order expectation, player 1 should also consult her third-order and fourth-order beliefs. Hence, there is an overlap in the causality diagram at order 3: one-fold belief in rationality and two-fold belief in rationality both impose restrictions on the third-order belief, and these restrictions may be in conflict.

[3]We slightly abuse the use of "k-th order expectation" here. Technically, a third-order expectation can be derived from the fifth-order expectation by taking the relevant marginal distribution. With third-order expectation in this context we specifically refer to $marg_{C_2} e_1^3 \in \Delta(C_2)$ where $e_1^3 \in \Delta(W_1^2 \times C_2)$ and by the fifth-order expectation we mean $marg_{C_2} e_1^5 \in \Delta(C_2)$ where $e_1^5 \in \Delta(W_1^4 \times C_2)$.

According to Theorem 2, there should also be expectation-based psychological games with utilities such that any kind of overlap in reasoning does not occur. In such games, as we will show in Section 6, iterated elimination of strictly dominated choices alone would *always* give exactly the choices one can make under common belief in rationality. We will do so by making use of causality diagrams.

5 Causality diagrams

Causality diagrams prove to be a useful analytical tool to think about the problem of overlap in reasoning which appears in psychological games. In order to formally capture the notion of a causality diagram, a discussion on elementary graph theory is in place. A *graph* Σ is a nonempty set of *vertices* V and a (possibly empty) set of *edges* E. In Figure 3 the vertices correspond to the choice and the orders of belief, whereas the edges are the arrows between the vertices which indicate in which orders a particular utility is variable. In a *directed graph*, the direction of the edge, also known as an *arc* in a directed graph, matters. In that case we speak of outgoing *arcs* if an arc leaves a vertex and ingoing arcs if an arc goes into a vertex. The amount of outgoing arcs is known as the outdegree, whereas the number of ingoing arcs is the indegree. All the vertices that some vertex x is joined with directly by an outgoing arc is known as the out-neighborhood, whereas all adjacent vertices that x is joined with via ingoing arcs is known as the in-neighborhood. Finally, there is the concept of a path. A *path* in a directed graph is a sequence of vertices that starts at the root, where the k-th element is joined with the $(k-1)$-th element by an ingoing arc. A vertex r is called a *root* if (a) that vertex has only outgoing arcs and (b) for all vertices in the graph that have ingoing arcs, there is a path from r to that vertex. Whenever we refer to a path in the remainder of this paper, we specifically mean a path that starts at the root. Moreover, we say we have a *divergence point* between two paths p^1 and p^2 in a rooted graph if $p^1 = (p, a, b, ...)$ and $p^2 = (p, a, c, ...)$ with p being a subpath of both p^1 and p^2 and $b \neq c$. The divergence point is then located at vertex a. The root itself can also function as the divergence point.

We can now formally define the concept of a causality diagram.

Definition 11. *Consider a game $G \in \mathcal{G}(N_1, N_2)$. The **causality diagram** $D_1(N_1, N_2)$ for player 1 in the game G is a rooted, directed graph $(\mathbb{N} \cup \{0\}, \mathcal{E})$ with the root being 0. The set of arcs $\mathcal{E}$ is as follows:*

- *For the **root** $r = 0$, establish an arc $(0, a_1)$ for every $a_1 \in N_1$;*
- *Inductively for every $k \geqslant 2$ do the following:*
  - *For every **even** a_{k-1}, establish an arc (a_{k-1}, a_k) with $a_k = a_{k-1} + b$ for every $b \in N_1$;*
  - *For every **odd** a_{k-1}, establish an arc (a_{k-1}, a_k) with $a_k = a_{k-1} + c$ for every $c \in N_2$.*

It is important to note that each player in a game has her own causality diagram, as players utilities may be variable in different orders. The paths in a causality diagram have a natural interpretation. Each path represents a chain of restrictions. For instance, in Figure 3 in order to ensure that some choice c_1 is optimal, the first-order and second-order expectation need to be restricted. If in addition player 1 wants to express 1-fold belief in rationality, restrictions on expectations of an even higher order are necessary. That is, given what player 1 expects player 2 to do, each choice in *that* conjecture can only be made optimal given its own appropriate restrictions on player 1's third-order and fifth-order expectations. And so on. In a traditional game, the causality diagram would look as in Figure 4. It is clear that in a traditional game there is only a single path on the causality diagram for each player. Compare this to the causality diagram in Figure 3. There, up to three orders, we can already distinguish between four different paths:

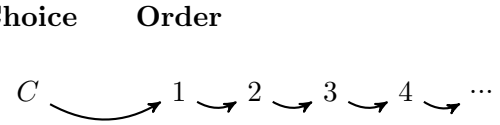


Figure 4: *Causality diagram of a player in a traditional game*

$(0, 1, 3, ...)$, $(0, 1, 5, ...)$, $(0, 2, 3, ...)$ and $(0, 2, 4, ...)$. The paths $(0, 1, 3, ...)$ and $(0, 2, 3, ...)$ also clearly display a common vertex after having diverged: vertex order 3. It is this overlap that can cause problems for the characterization of rational choices under common belief in rationality by IESDC.

We say that two paths in a directed graph are *pairwise vertex-disjoint* starting at a particular vertex a if they do not have any vertices in common after this vertex a. This leads us to define the following concept.

Definition 12. *A causality diagram is overlap-free if all pairs of paths are pairwise vertex-disjoint after the respective divergence point.*

If we take the interpretation that a path represents a chain of utility-relevant restrictions, then if two paths have a vertex b in common after a divergence point, it follows that two paths lead to two different restrictions on the same set of higher-order beliefs. Of course, these restrictions may clash. If paths are vertex-disjoint however, then the set of higher-order expectations for a particular order is never restricted from multiple angles.

The combinations of variable orders N_1 and N_2 that induce an overlap-free causality diagram are actually identifiable. These correspond to the three cases listed in Theorem 2.

Lemma 2. *The causality diagram $D_1(N_1, N_2)$ for player 1 in a game $G \in \mathcal{G}(N_1, N_2)$ is overlap-free if and only if one of the following is the case:*

(i) $N_1 = \{x\}$ *and* $N_2 = \{y\}$*;*

(ii) $N_1 = \{x\}$ *with* x *even;*

(iii) N_1 *only consists of odd orders and* $N_2 = \{z\}$ *with* z *even such that there is no pair* $x, y \in N_1$ *and no* $n \in \mathbb{N}$ *where* $x + n \cdot z = y$*.*

Proof. We start off by proving the "if"-direction. We do so for each of the three cases separately. Canonical causality diagrams that represent each of the cases are depicted in Figure 5.
$\Leftarrow$:
(i) If $N_1 = \{x\}$ and $N_2 = \{y\}$, then the cardinality of both sets of orders is one. This implies that in the resulting causality diagram every vertex has an outdegree of at most one. Then it follows there is also a unique path in the causality diagram. By definition the causality diagram is then overlap-free, as there is no second path present to have overlap with.

(ii) Clearly, there is a unique path in player 1's causality diagram, containing only even numbers. Hence vertex-disjointness is guaranteed in this case and therefore the causality diagram is overlap-free. This is illustrated in Figure 5b.

(iii) The root may be at the start of multiple paths, as $|N_1| \geqslant 1$. As the out-neighborhood of the root is determined by N_1, the root is only connected to $x \in N_1$, each of which is odd. Each odd vertex's out-neighborhood is determined by $N_2 = \{z\}$. Thus each odd vertex is connected to a different odd vertex, as z is even. Then take two paths in player 1's causality diagram:

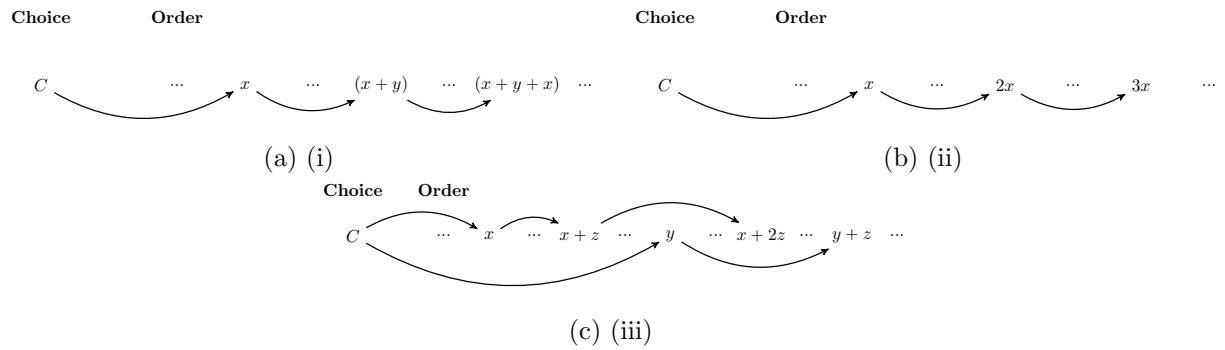


Figure 5: *Canonical causality diagram for cases (i), (ii) and (iii) in Lemma 2*

$(0, x, x+z, ..., x+g\cdot z, ...)$ and $(0, x', x'+z, ..., x'+h\cdot z, ...)$ with $x, x' \in N_1$ and $x \neq x'$. These two paths must be vertex-disjoint. First, note that the divergence point of the two paths is the root. Now assume they do have a vertex in common: let $x + g\cdot z = x' + h\cdot z$. Then $x - x' = (h-g)\cdot z$. Assume without loss of generality that $h > g$. However, this violates the condition that there is no pair $x, x' \in N_1$ such that $x' + n\cdot z = x$ with $n = h - g$. Hence it must be the case that all paths in player 1's causality diagram are pairwise vertex-disjoint and therefore the causality diagram is overlap-free.

For the"only-if" direction, we show that if conditions (i), (ii) and (iii) do not hold, then the causality diagram of player 1 has an overlap. In total there are six scenarios under which none of the three cases listed in Lemma 2 apply: (i) N_1 contains two even orders; (ii) N_1 contains an even and an odd order, whereas N_2 contains an even order; (iii) N_1 contains an even and an odd order, whereas N_2 contains an odd order; (iv) N_1 contains an odd x and an odd y and N_2 contains an even z such that $x = y + n\cdot z$ for some $n \in \mathbb{N}$; (v) N_1 contains two odd orders and N_2 contains also an odd order; and (vi) N_1 contains an odd order and N_2 contains two arbitrary orders. We go by all these cases one-by-one.

$\Rightarrow$:
(i) Consider $\{x, y\} \subseteq N_1$ with both x and y even. Then there exist the following two paths in the causality diagram $D_1(N_1, N_2)$: $(0, x, 2x, ...)$ and $(0, y, 2y, ...)$. These paths share a common vertex in $x \cdot y$. Hence the two paths are not vertex-disjoint after their divergence-point and thus the causality diagram $D_1(N_1, N_2)$ for player 1 is not overlap-free.
(ii) Consider $\{x, y\} \subseteq N_1$ with x even and y odd and an even $z \in N_2$. We have a path $(0, y, y + z, y + 2\cdot z, ..., y + x\cdot z, ...)$. There is also the path $(0, x, 2x, ..., z\cdot x, z\cdot x + y, ...)$. These two paths have vertex $z\cdot x + y$ in common after the point of divergence, being the root. Hence the causality diagram is not overlap-free.
(iii) Now consider $\{x, y\} \subseteq N_1$ with x even and y odd and an odd $z \in N_2$. We have a path $(0, y, y + z, y + (y + z), (y + z) + (y + z), ..., x(y + z), ...)$. There exists also the path $(0, x, 2x, ..., (y + z)x, ...)$. These two paths share a vertex in $x(y + z)$ after the divergence point, being the root. Hence the causality diagram is not overlap free.
(iv) Consider $\{x, y\} \subseteq N_1$ with both x and y odd and an even $z \in N_2$ such that $x = y + n\cdot z$ for some $n \in \mathbb{N}$. There is a path $(0, y, y + z, y + 2\cdot z..., y + n\cdot z, ...)$. We assumed that $x = y + n\cdot z$. We then have two paths: $(0, x, ...)$ and $(0, y, y + z, y + 2\cdot z, ..., y + n\cdot z, ...)$. These two paths share a vertex in $y + n\cdot z = x$. Hence the causality-diagram for player 1 is not overlap-free.

(v) Consider $\{x, y\} \subseteq N_1$ with both x and y odd and an odd $z \in N_2$. Then the paths $(0, x, x + z, x + z + y, ...)$ and $(0, y, y + z, y + z + x, ...)$ share a vertex after the point of divergence in $x + z + y$. Hence the resulting causality diagram for player 1 contains paths that are not vertex-disjoint after the divergence point and thus is not overlap-free.
(vi) Consider $\{x\} \subseteq N_1$ with x odd and consider two orders $\{y, z\} \subseteq N_2$. The root of the causality diagram of player 1 is then joined to the odd vertex x. This vertex x can be considered as the root of its own subgraph D_2^x. Then, if N_2 contains two even orders, the scenario (i) applies to subgraph D_2^x. Hence, under such a scenario, D_2^x would not be overlap-free. If N_2 contains an even and an odd order, the scenario of (iii) applies to D_2^x. Then also now, D_2^x is not overlap-free. If N_2 contains two odd orders, then the scenario of (v) applies. Also then, D_2^x is not overlap-free. Since the root of player 1's causality diagram is connected to vertex x, the paths that are not pairwise vertex-disjoint in D_2^x are subpaths in player 1's causality diagram. Hence there also exist paths in player 1's causality diagram that are not vertex-disjoint after a point of divergence. Therefore, player 1's causality diagram is also not overlap-free.

Hence, we have shown that if none of the three cases listed in Lemma 2 apply, the causality diagram cannot be overlap-free. This ends the proof for Lemma 2.

□

In the following section, we will provide and discuss the proof for Theorem 2.

6 Proof of Theorem 2

In order to prove Theorem 2, we will split it up into two seperate lemmas: Lemmas 3 and 4. We will prove each of those in turn. For Lemma 4 we provide sketches of the proof in this section. The full proof can be found in the Appendix accompanying this paper.

Lemma 3. *Consider a family of games $\mathcal{G}(N_1, N_2)$. If for every game in $\mathcal{G}(N_1, N_2)$, each choice that survives the IESDC-procedure for player 1 is also a rational choice under common belief in rationality, then the causality diagram of player 1 is overlap-free.*

Proof. Suppose that the causality diagram $D_1(N_1, N_2)$ of player 1 is not overlap-free. We will construct a game G^* in $\mathcal{G}(N_1, N_2)$ such that not every choice for player 1 that survives the IESDC-procedure is rational under common belief in rationality.

We will do this in the following way. First, assume (N_1, N_2) induces a causality diagram for player 1 where the root is already a divergence point for two paths that overlap. So we have a path $(0, a_1, a_2, ..., a_{n-1}, a_n = s, ...)$ and a path $(0, z_1, z_2, ..., z_{m-1}, z_m = s, ...)$. Subsequently we construct a game G^* that has some particular properties. First, we construct G^* such that each choice in C_1 and each choice in C_2 survives the IESDC-procedure. So each choice is optimal for *some* belief hierarchy. Then we construct the game G^* further such that $\bar{c}_1 \in C_1$ is only optimal for a belief hierarchy whose a_1-th order expectation assigns probability one to choice $c[a_1]$ (in C_1^∞ if a_1 is even and in C_2^∞ if a_1 is odd) *and* whose z_1-th order expectation assigns probability one to choice $d[z_1]$ (also in C_1^∞ if z_1 is even and in C_2^∞ if a_1 is odd). Choice $c[a_1]$ in turn is only optimal for a $(a_2 - a_1)$-th order expectation assigning probability one to choice $c[a_2]$. And so on, up until we arrive at choice $c[a_n]$ for the s-th order expectation. We can do the same for the second subpath $(0, z_1, z_2, ..., z_m = s)$, which then ends up with a choice $d[z_m]$ for the s-th order expectation. We construct G^* such that $c[a_n] \neq d[z_m]$, and hence $\bar{c}_1$ cannot be optimal while expressing common

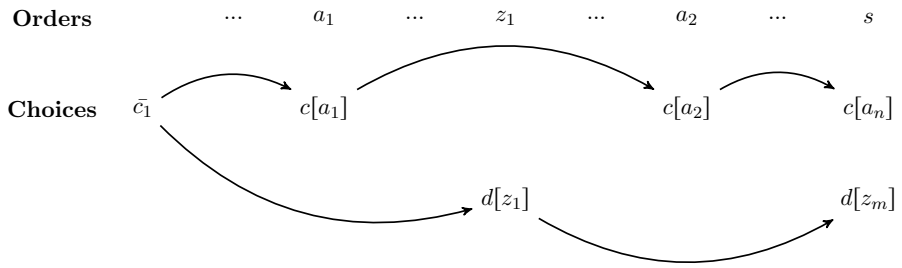


Figure 6: Example Proof Lemma 3, (part of) beliefs diagram

belief in rationality.

Now, let
$C_1 := \{c[a_k] : a_k \text{ is even and } k \in \{1, ..., n\}\} \cup \{d[z_k] : z_k \text{ is even and } k \in \{1, ..., n\}\} \cup \{\bar{c}_1\} \cup \{x\}$, and
$C_2 := \{c[a_k] : a_k \text{ is odd and } k \in \{1, ..., n\}\} \cup \{d[z_k] : z_k \text{ is odd and } k \in \{1, ..., n\}\} \cup \{y\}$. We assume here that all these choices are different. Note that for order $z_m = a_n = s$, there are two different choices: $c[a_n]$ and $d[z_m]$. Let the choices have the following properties.

1. Choice x always yields a utility of 1 for player 1, under any belief hierarchy b_1;
2. Choice y always yields a utility of 1 for player 2, under any belief hierarchy b_2;
3. Choice $\bar{c}_1$ is only optimal for a higher-order expectation which assigns in the a_1-th component probability one to $c[a_1]$ and in the z_1-th component probability one to $d[z_1]$. Only in that case choice $\bar{c}_1$ leads to a utility of 1. In all other cases, utility is 0;
4. Each $c[a_{k-1}] \in C_1$ for $k \in \{2, ..., n\}$ is such that $u_1(c[a_{k-1}], b_1) = 1$ for any $b_1 \in B_1$ where the $(a_k - a_{k-1})$-th order expectation assigns probability one to $c[a_k]$. If $b_1 \in B_1$ is such that the $(a_k - a_{k-1})$-th order expectation assigns probability one to any choice $c \neq c[a_k]$, then the utility will be 0. Each $d[z_{k-1}] \in C_1$ for $k \in \{2, ..., n\}$ is such that $u_1(d[z_{k-1}], b_1) = 1$ for any $b_1 \in B_1$ where the $(z_k - z_{k-1})$-th order expectation assigns probability one to $d[z_k]$. If $b_1 \in B_1$ is such that the $(z_k - z_{k-1})$-th order expectation assigns probability one to any choice $d \neq d[z_k]$, then the utility will be 0. We do exactly the same for each $c[a_{k-1}], d[z_{k-1}] \in C_2$ for $k \in \{2, ..., n\}$.
5. For simplicity assume $c[a_n]$ and $d[z_m]$ *always* lead to a utility of 1.

Note that *each* choice in C_1 and C_2 leads to a utility of 1 under some extreme higher-order expectation. As a result, no choice is strictly dominated and hence each choice survives the IESDC-procedure.

We will now show that choice $\bar{c}_1$ in the game G^* as constructed before cannot be optimal under a belief hierarchy expressing common belief in rationality, even though it survives the IESDC-procedure. We will do so by arguing that $\bar{c}_1$ cannot be optimal for a belief hierarchy that simultaneously expresses up to a_{n-1}-fold and up z_{m-1}-fold belief in rationality. That is, the latter two events will require conflicting restrictions on the s-th order expectation. Figure 6 helps in illustrating this point by depicting (part of) a beliefs diagram of some game G^*.

In game G^*, choice $\bar{c}_1$ is optimal only if the a_1-th order expectation of player 1 assigns probability one to choice $c[a_1]$. Namely, only then the utility for player 1 is equal to 1. For any $k \in \{2, ..., n\}$,

choice $c[a_{k-1}]$ is only optimal if the $(a_k - a_{k-1})$-th order expectation (of player 1 if a_{k-1} is even, otherwise of player 2) assigns probability one to choice $c[a_k]$. Thus we obtain a chain of restrictions. In the example of Figure 6 this would correspond to the upper path $(0, a_1, a_2, s)$. In order for player 1 to be able to rationally choose $\bar{c}_1$ under a belief hierarchy expressing up to a_{n-1}-fold belief in rationality, the s-th order expectation should thus assign probability one to choice $c[s]$.

Another requirement for choice $\bar{c}_1$ in game G^* to be optimal is that the z_1-th order expectation of player 1 assigns probability one to choice $d[z_1]$. Namely, only under such a condition can the utility of player 1 be equal to 1. For any $k \in \{2, ..., n\}$, choice $d[z_{k-1}]$ is only optimal if the $(z_k - z_{k-1})$-th order expectation (of player 1 if z_{k-1} is even, otherwise of player 2) assigns probability one to choice $d[z_k]$. In Figure 6 this chain of restrictions would for instance correspond to the lower path $(0, z_1, s)$. In order for player 1 to be able to rationally choose $\bar{c}_1$ under a belief hierarchy expressing up to z_{m-1}-fold belief in rationality, the s-th order expectation should assign probability one to choice $d[s]$.

We constructed G^* such that $c[s] \neq d[s]$. But then, $\bar{c_1}$ cannot be optimal for a belief hierarchy that expresses both up to a_{n-1}-fold belief in rationality and up to z_{m-1}-fold belief in rationality. Then $\bar{c_1}$ also cannot be optimal for a belief hierarchy expressing common belief in rationality.

Now, we initially assumed that the point of divergence for our two paths was at the root. Not for every causality diagram with overlap this is possible. However, it should certainly be possible to occur within one arc-distance of the root.

Claim 1. *Consider a causality diagram $D_1(N_1, N_2)$ for player 1 that is not overlap-free. Then there exist two paths with overlap that either (a) have a point of divergence at the root, or (b) that have a point of divergence at an odd order $a \in N_1$.*

Proof of claim. To prove this, we can point to our proof construction for the "only if"-part of the proof for Lemma 2. In this proof, we provided an exhaustive list of six scenarios. In the first five scenarios listed for that proof we were able to construct two paths with overlap that had a divergence point at the root of the causality diagram. So for these scenarios case (a) of the claim is satisfied. For the final scenario we noted that the root had an outgoing arc to an odd vertex $a \in N_1$. This odd vertex was the root of its own subgraph, which always could be categorized under scenario (i), (iii) or (v). As such, in this subgraph there also existed overlapping paths that had their point of divergence at the root. Hence, for this scenario case (b) is satisfied. This completes the proof of this claim.

So let us now consider the case that the first point of divergence in the causality diagram occurs at the odd order $a \in N_1$. Let us have two paths $(0, a, a_1, a_2, ..., a_{n-1}, a_n = s, ...)$ and $(0, a, z_1, z_2, z_{m-1}, z_m = s, ...)$. Then we can simply construct the game G^* as follows. Let
$C_1 := \{c[a_k] : a_k \text{ is even and } k \in \{1, ..., n\}\} \cup \{d[z_k] : z_k \text{ is even and } k \in \{1, ..., n\}\} \cup \{\bar{c}_1\} \cup \{x\}$, and
$C_2 := \{c[a_k] : a_k \text{ is odd and } k \in \{1, ..., n\}\} \cup \{d[z_k] : z_k \text{ is odd and } k \in \{1, ..., n\}\} \cup \{\bar{c}_2\} \cup \{y\}$. Again, we assume here that all these choices are different. Let the choices have the following properties.

1. Choice x always yields a utility of 1 for player 1, under any belief hierarchy b_1;
2. Choice y always yields a utility of 1 for player 2, under any belief hierarchy b_2;
3. Choice $\bar{c}_2$ is only optimal for a higher-order expectation which assigns in the $(a_1 - a)$-th component probability one to $c[a_1]$ and in the $(z_1 - a)$-th component probability one to $d[z_1]$. Only in that case choice $\bar{c}_1$ leads to a utility of 1. In all extreme other cases, utility is 0;
4. Each $c[a_{k-1}] \in C_1$ for $k \in \{2, ..., n\}$ is such that $u_1(c[a_{k-1}], b_1) = 1$ for any $b_1 \in B_1$ where the $(a_k - a_{k-1})$-th order expectation assigns probability one to $c[a_k]$. If $b_1 \in B_1$ is such that

the $(a_k - a_{k-1})$-th order expectation assigns probability one to any choice $c \neq c[a_k]$, then the utility will be 0. Each $d[z_{k-1}] \in C_1$ for $k \in \{2, ..., n\}$ is such that $u_1(d[z_{k-1}], b_1) = 1$ for any $b_1 \in B_1$ where the $(z_k - z_{k-1})$-th order expectation assigns probability one to $d[z_k]$. If $b_1 \in B_1$ is such that the $(z_k - z_{k-1})$-th order expectation assigns probability one to any choice $d \neq d[z_k]$, then the utility will be 0. We do exactly the same for each $c[a_{k-1}], d[z_{k-1}] \in C_2$ for $k \in \{2, ..., n\}$.

5. For simplicity assume $c[a_n]$ and $d[z_m]$ *always* lead to a utility of 1;
6. Choice $\bar{c}_1$ is only optimal for a higher-order expectation which assigns in the a-th component probability one to $\bar{c}_2$. Only in that case choice $\bar{c}_1$ leads to a utility of 1. In all extreme other cases, utility is 0.

Note that game G^* from the perspective of player 2 is exactly as it was before from the perspective of player 1. It follows then that choice $\bar{c}_2$ cannot be optimal for player 2 under a belief hierarchy expressing common belief in rationality. In the new version of G^* we added an extra choice for player 1: choice $\bar{c}_1$ is only optimal under an a-th order expectation that assigns probability one to choice $\bar{c}_2$ for player 2. Then it follows that choice $\bar{c}_1$ is also never optimal given a belief hierarchy expressing common belief in rationality.

Since we took an arbitrary combinations (N_1, N_2) that leads to a causality diagram for player 1 with overlap, it follows that for each family of games $\mathcal{G}(N_1, N_2)$ we can construct a game G^* as we did here. □

From Lemma 3 we can conclude that if the causality diagram is not overlap-free, there always exist accompanying psychological games in which the IESDC-procedure does not characterize those choices that are rational under common belief in rationality. Next we will show that if the causality diagram *is* overlap-free, then this exact characterization *does* always occur.

Lemma 4. *Consider a family of games $\mathcal{G}(N_1, N_2)$. If the causality diagram of player 1 is overlap-free, then for every game in $\mathcal{G}(N_1, N_2)$, each choice that survives the IESDC-procedure for player 1 is also a rational choice under common belief in rationality.*

The actual proof can be found in the appendix. Here we provide an overview of the proof and an explanation by means of examples.

The outline of this proof is as follows. We consider three different scenarios, which together exactly cover all three cases from Theorem 2. Scenario (i) corresponds to the scenario that N_1 contains a single even order, scenario (ii) correspond to the case that N_1 and N_2 both consists of a single odd order, and scenario (iii) corresponds to case (iii) of Theorem 2. Note here that scenario (iii) covers the subcase (i) of Theorem 2 where N_1 contains a single odd order and N_2 contains a single even order.

The proof is constructive. For each of the scenarios listed above, we will take some arbitrary choice $c_1 \in C_1^\infty$ that survives the IESDC-procedure. The goal is to construct a belief hierarchy that expresses common belief in rationality and is such that it optimizes choice c_1. We do so by making use of finite epistemic models where types represent belief hierarchies.

Definition 13 (Epistemic model in a static psychological game)**.**
*Consider a psychological game G. An **epistemic model** $M = (T_i, b_i)_{i \in I}$ for G specifies for every player i a finite set T_i of possible types. Moreover, for every player i and every type $t_i \in T_i$ the epistemic model specifies a probability distribution $b_i[t_i]$ over the set of the opponent's choice-type combinations $C_j \times T_j$. The probability distribution $b_i[t_i]$ represents the belief player i has about the choice-type combinations of her opponent.*

Each type represents a probability distribution over the opponent's choice-type combination. Each types thus already induces a probability distribution over the opponent's choices. Hence we can retrieve first-order beliefs from types in an epistemic model. Since a type also represents a probability distribution over the opponent's types, we can also retrieve a probability distribution over the opponent's first-order beliefs. As such, types also capture second-order beliefs. In a similar fashion we can retrieve from a type in an epistemic model third-order beliefs, fourth-order beliefs, and so on.

The proof of Lemma 4 consists of two steps. In Step 1, we first fix an arbitrary choice $c_1 \in C_1^\infty$ for player 1 that survives the IESDC-procedure. For each of the three scenarios, we construct a finite epistemic model with a type $t_1^{c_1}$ that optimizes choice c_1. Moreover, we construct this epistemic model such that for each order of belief k that is on a path in player 1's causality diagram, the type $t_1^{c_1}$ expresses k-fold belief in rationality. We call this on-path belief in rationality.

Definition 14. *Consider a game in $\mathcal{G}(N_1, N_2)$ and the causality diagram $D_1(N_1, N_2)$ for player 1. Consider a belief hierarchy b_1 for player 1. We say b_1 expresses **on-path belief in rationality** if it expresses k-fold belief in rationality for every $k \geqslant 1$ that is on some path in the causality diagram.*

Then, in Step 2, we transform the epistemic model created in Step 1. We ensure that for all the remaining orders of belief l, type $t_1^{c_1}$ expresses also l-fold belief in rationality. Then type $t_1^{c_1}$ will also express common belief in rationality. A formal proof can be found in the appendix. Here, we sketch the proof by means of examples for the three scenarios.

Example Scenario (i) First consider scenario (i). This corresponds to case (ii) in Theorem 2. Let $C_1 = \{A, B, C\}$, $C_2 = \{D, E\}$, $N_1 = \{4\}$ and let $N_2 = \{1, 2\}$. Consider the game in Table 5. Each choice for player 1 and player 2 in this game is not strictly dominated, and hence

Table 5: *Illustration of Proof scenario* (i)

Player 1's extreme fourth-order expectations

	A	B	C
A	2	2	0
B	3	0	2
C	0	3	1

Player 1's utilities

Player 2's combinations of extreme first-order and second-order expectations

	(A,D)	(A,E)	(B,D)	(B,E)	(C,D)	(C,E)
D	0	3	0	1	2	0
E	1	1	1	0	1	2

Player 2's utilities

$C_1^\infty = \{A, B, C\}$, $C_2^\infty = \{D, E\}$.

First, for A, B and C we will fix fourth-order expectations that will optimize each of these choices in turn. For A, we can take $b_1^A \in \Delta(C_1^\infty)$ with $b_1^A = 0.5A + 0.5B$. For B, we can take $b_1^B = C$ and for C we can take $b_1^C = B$.

Next, for each choice in $c_1 \in C_1^\infty$, we will construct a type $t_1^{c_1}$ for player 1. Each such type assigns probability one to a type $t_2^{c_1,1}$ for player 2, which on its turn assigns probability one to a type $t_1^{c_1,2}$ for player 1. Type $t_1^{c_1,2}$ on its turn assigns probability one to type $t_2^{c_1,3}$ for player 2. Type $t_2^{A,3}$ is such that $b_2[t_2^{A,3}] = 0.5(A, t_1^A) + 0.5(B, t_1^B)$; type $t_2^{B,3}$ is such that $b_2[t_2^{B,3}] = (C, t_1^C)$;

type $t_2^{C,3}$ is such that $b_2[t_2^{C,3}] = (B, t_1^B)$. The probability distributions over choices in the induced beliefs here are thus equal to the fourth-order expectations we fixed before to optimize each choice. We couple the choices assigned positive probability to in these induced beliefs with the respective types we started with. The resulting, partial epistemic model is illustrated via the beliefs diagram in Figure 7a. We explictely mention that it is partial, since some types induce beliefs that do not yet specify beliefs over choices.

Next, we continue with Step 2 of the proof for this scenario. This involves filling in all blank spaces in our beliefs diagram. We do so in an iterative way. First, fill in random choices in each blank space. This is illustrated in a beliefs diagram in Figure 7b, by taking all components that have a superscript 0.

Next, take each right-most matrix. For instance, take the upper-right matrix in the diagram of Figure 7b. In this order of belief player 2 expects with probability 0.5 player 1 to choose A while expecting player 1 to believe that player 2 plays E^0. With the remaining probability of 0.5 player 2 expects player 1 to choose B while expecting player 1 to believe that player 2 plays D^0. For such a second-order belief, choice D is optimal. Hence we fill in D^1 in the upper-right matrix. Then, take the upper-middle matrix. Here player 1 in her fourth-order expectation assigns 0.5 to choice C^0 and probability 0.5 to choice B^0. Then choice C is optimal. Hence we fill in C^1 in the upper-middle matrix. Lastly, take the left-upper matrix. Here now player 2 expects player 1 to choose C^1 while expecting player 1 believes player 2 will choose D^1. Thus we get a second-order expectation of (C, D). Given (C, D), choice D is optimal for player 2. Hence, in the left-upper matrix we list choice D^1. We do this for every sequence of matrices in the diagram.

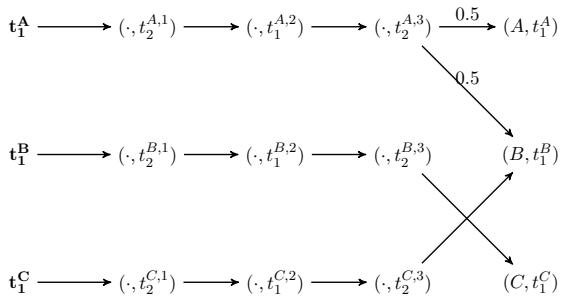
Next, in Iteration 2, we do a similar thing, leading to the choices with superscript 2 in the various matrices.

After a while we see a pattern emerge. We see that iterations 2 and 3 correspond to iterations 4 and 5 respectively. So the pattern repeats itself always after two iterations. Also, we have from iteration 1 onwards that each type expresses belief in the opponent's rationality by construction. Taken together, two recurring iterations from iteration 1 onwards are therefore sufficient to characterize a finite, epistemic model in which each type expresses common belief in rationality. In this case these are Iterations 2 and 3. We loop these iterations indefinitely. In this way, we construct the epistemic model $\mathcal{M}^*$ as is depicted in Figure 7c. All beliefs in the second column of the table in this figure correspond to beliefs generated by Iteration 3 of Step 2. All beliefs in the last column of the table correspond to beliefs generated by Iteration 2 of Step 2.

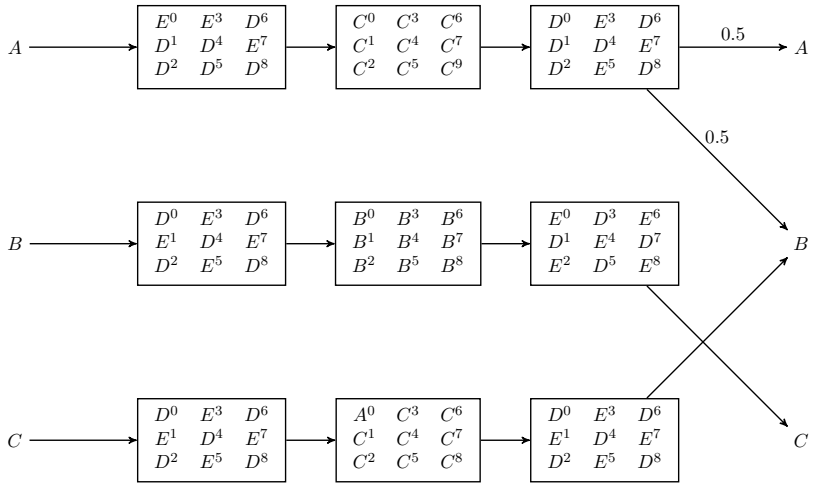
One may verify that in $\mathcal{M}^*$ each type expresses common belief in rationality. This includes types t_1^A, t_1^B and t_1^C . By construction of Step 1 these types optimize choices A, B and C respectively. Hence, for each choice that survives the IESDC-procedure we have managed to construct a type expressing common belief in rationality, such that that choice is also still optimal.

Example Scenario (ii) Consider scenario (ii). This corresponds to case (i) in Theorem 2 where for both players the orders are odd. Let, $C_1 = \{A, B, C\}$, $C_2 = \{D, E\}$, $N_1 = \{3\}$ and let $N_2 = \{1\}$. Consider the game in Table 6. Note that each choice for player 1 is not strictly dominated. Similarly, for player 2, both choices D and E are not strictly dominated.

First, for A, B and C we will fix third-order expectations that we will optimize each of these choices in turn. For A, we can take $b_1^A \in \Delta(C_1^\infty)$ with $b_1^A = 0.5D + 0.5E$. For B, we can take $b_1^B = E$ and for C we can take $b_1^C = D$. Similary, for choice D of player 2 we can take $b_2^D = 0.6A + 0.4B$ and for choice E we can take $b_2^E = C$.



(a) Step 1 scenario (i)



(b) Beliefs diagram Step 2

Types player 1	$T_1 = \{t_1^A, t_1^B, t_1^C, t_1^{A,2}, t_1^{B,2}, t_1^{C,2}, t_1^{A\prime}, t_1^{B\prime}, t_1^{C\prime}, t_1^{A,2\prime}, t_1^{B,2\prime}, t_1^{C,2\prime}\}$						
Types player 2	$T_2 = \{t_2^{A,1}, t_2^{B,1}, t_2^{C,1}, t_2^{A,3}, t_2^{B,3}, t_2^{C,3}, t_2^{A,1\prime}, t_2^{B,1\prime}, t_2^{C,1\prime}, t_2^{A,3\prime}, t_2^{B,3\prime}, t_2^{C,3\prime}\}$						
Player 1's beliefs	$b_1[t_1^A]$	=	$(E, t_2^{A,1})$		$b_1[t_1^{A\prime}]$	=	$(D, t_2^{A,1\prime})$
	$b_1[t_1^B]$	=	$(E, t_2^{B,1})$		$b_1[t_1^{B\prime}]$	=	$(D, t_2^{B,1\prime})$
	$b_1[t_1^C]$	=	$(E, t_2^{C,1})$		$b_1[t_1^{C\prime}]$	=	$(D, t_2^{C,1\prime})$
	$b_1[t_1^{A,2}]$	=	$(E, t_2^{A,3})$		$b_1[t_1^{A,2\prime}]$	=	$(D, t_2^{A,3\prime})$
	$b_1[t_1^{B,2}]$	=	$(D, t_2^{B,3})$		$b_1[t_1^{B,2\prime}]$	=	$(E, t_2^{B,3\prime})$
	$b_1[t_1^{C,2}]$	=	$(E, t_2^{C,3})$		$b_1[t_1^{C,2\prime}]$	=	$(D, t_2^{C,3\prime})$
Player 2's beliefs	$b_2[t_2^{A,1}]$	=	$(C, t_1^{A,2})$		$b_2[t_2^{A,1\prime}]$	=	$(C, t_1^{A,2\prime})$
	$b_2[t_2^{B,1}]$	=	$(B, t_1^{B,2})$		$b_2[t_2^{B,1\prime}]$	=	$(B, t_1^{B,2\prime})$
	$b_2[t_2^{C,1}]$	=	$(C, t_1^{C,2})$		$b_2[t_2^{C,1\prime}]$	=	$(C, t_1^{C,2\prime})$
	$b_2[t_2^{A,3}]$	=	$0.5(A, t_1^{A\prime}) + 0.5(B, t_1^{B\prime})$		$b_2[t_2^{A,3\prime}]$	=	$0.5(A, t_1^A) + 0.5(B, t_1^B)$
	$b_2[t_2^{B,3}]$	=	$(C, t_1^{C\prime})$		$b_2[t_2^{B,3\prime}]$	=	(C, t_1^C)
	$b_2[t_2^{C,3}]$	=	$(B, t_1^{B\prime})$		$b_2[t_2^{C,3\prime}]$	=	(B, t_1^B)

(c) Epistemic model scenario (i)

Figure 7: Illustration proof Scenario (i)

Table 6: *Illustration of Proof scenario* (ii)

Player 1's extreme third-order expectations

	D	E
A	2	2
B	0	3
C	3	0

Player 1's utilities

Player 2's extreme first-order expectations

	A	B	C
D	6	6	6
E	8	0	8

Player 2's utilities

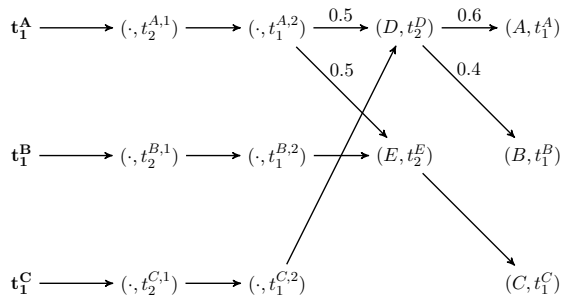
Next, for each choice $c \in C_1^\infty$, we can construct a type t_1^c. Similarly, for each choice $c \in C_2^\infty$, we can construct a type t_2^c. Each type t_1^c assigns probability one to a type $t_2^{c,1}$, which on its turn induces a belief that assigns probability one to a type $t_1^{c,2}$. Type $t_1^{A,2}$ is such that $b_1[t_1^{A,2}] = 0.5(D, t_2^D) + 0.5(E, t_2^E)$; type $t_1^{B,2}$ is such that $b_1[t_1^{B,2}] = (E, t_2^E)$; type $t_1^{C,2}$ is such that $b_1[t_1^{C,2}] = (D, t_2^D)$. Similarly, we can let type t_2^D be such that $b_2[t_2^D] = 0.6(A, t_1^A) + 0.4(B, t_1^B)$ and type t_2^E be such that $b_2[t_2^E] = (C, t_1^C)$ The resulting, partial epistemic model is illustrated via the beliefs diagram in Figure 8a.

We then use exactly the same construction method as we did for situation (i) for Step 2. First, for the blank spaces we fill in random choices. These choices have superscript 0 and can be seen in Figure 8b. Next, for each sequence of matrices, we will select optimal choices in a backward fashion. For instance, take the right-upper matrix. This order of belief relates to choices of player 1. We know that player 1's utility is variable in order 3. According to the diagram, player 1's third-order expectation reasoned from this matrix is $0.5 \cdot 0.6(D, A, E^0) + 0.5 \cdot 0.4(D, B, D^0) + 0.5(E, C, D^0)$. Summarized, this yields $0.3 \cdot E^0 + 0.7 \cdot D^0$ as a third-order expectation. Given this third-order expectation, we have that choice C is optimal. Hence we list choice C^1 next in this matrix. Similarly, take the left-upper matrix. We know player 2's utility is variable in order 1. According to the diagram, player 2's first-order expectation reasoned from this point is one that assigns probability one to choice C^1. Given this first-order expectation, choice E is optimal. Hence we list E^1. We do this same backward construction for the lower sequence of matrices in the beliefs diagram as well. Now, for the next iteration of Step 2, we do something similar, leading to the choices with superscript 2 in the various matrices.

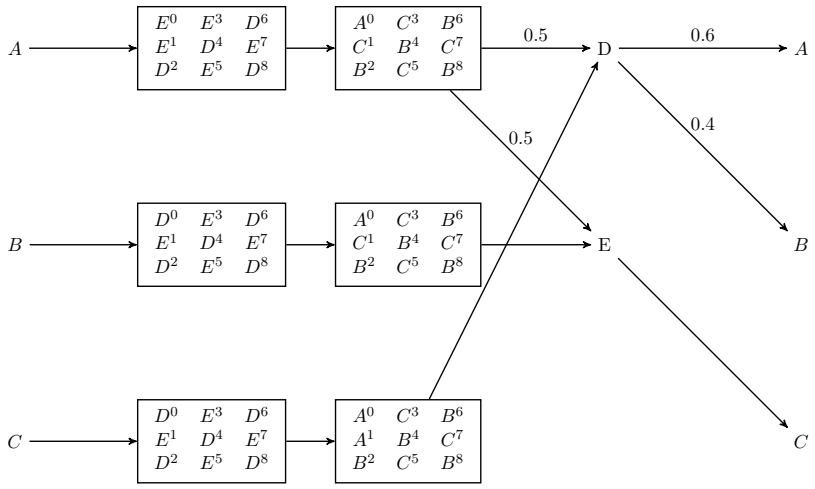
Just like in scenario (i), we continue this process iteratively. We can note that iterations 4 and 5 yield the same choices as iterations 2 and 3. From Iteration 1 onwards we have that each type constructed expresses belief in the opponent's rationality. Hence, by looping Iteration 2 and 3 we can construct a finite, epistemic model where each type expresses common belief in rationality. Moreover, this would also include types that optimize each choice that survives the IESDC-procedure for player 1 for the same reason as in scenario (i). The resulting epistemic model $\mathcal{M}^*$ is found in Figure 8c. All beliefs in the second column of the table in this figure correspond to beliefs generated by Iteration 3 of Step 2. All beliefs in the last column of the table correspond to beliefs generated by Iteration 2 of Step 2.

Scenario (iii)

Consider scenario (iii). This corresponds to case (iii) in Theorem 2. Take $N_1 = \{1, 7\}$ and $N_2 = \{4\}$. Now, consider the game as depicted in Table 7, with $C_1 = \{a, b\}$ and $C_2 = \{c, d, e, f\}$. Note that no choice for player 2 is strictly dominated. No choice for player 1 is strictly dominated either. Thus $C_1 = C_1^\infty$ and $C_2 = C_2^\infty$. Also note that choice a for player 1 is only optimal for



(a) Step 1 scenario (*ii*)



(b) Beliefs diagram iteration Step 2

Types player 1	$T_1 = \{t_1^A, t_1^B, t_1^C, t_1^{A,2}, t_1^{B,2}, t_1^{C,2}, t_1^{A\prime}, t_1^{B\prime}, t_1^{C\prime}, t_1^{A,2\prime}, t_1^{B,2\prime}, t_1^{C,2\prime}\}$					
Types player 2	$T_2 = \{t_2^{A,1}, t_2^{B,1}, t_2^{C,1}, t_2^D, t_2^E, t_2^{A,1\prime}, t_2^{B,1\prime}, t_2^{C,1\prime}, t_2^{D\prime}, t_2^{E\prime}\}$					
Player 1's beliefs	$b_1[t_1^A]$	=	$(E, t_2^{A,1})$	$b_1[t_1^{A\prime}]$	=	$(D, t_2^{A,1\prime})$
	$b_1[t_1^B]$	=	$(E, t_2^{B,1})$	$b_1[t_1^{B\prime}]$	=	$(D, t_2^{B,1\prime})$
	$b_1[t_1^C]$	=	$(E, t_2^{C,1})$	$b_1[t_1^{C\prime}]$	=	$(D, t_2^{C,1\prime})$
	$b_1[t_1^{A,2}]$	=	$0.5(D, t_2^D) + 0.5(E, t_2^E)$	$b_1[t_1^{A,2\prime}]$	=	$0.5(D, t_2^{D\prime}) + 0.5(E, t_2^{E\prime})$
	$b_1[t_1^{B,2}]$	=	(E, t_2^E)	$b_1[t_1^{B,2\prime}]$	=	$(E, t_2^{E\prime})$
	$b_1[t_1^{C,2}]$	=	(D, t_2^D)	$b_1[t_1^{C,2\prime}]$	=	$(D, t_2^{D\prime})$
Player 2's beliefs	$b_2[t_2^{A,1}]$	=	$(C, t_1^{A,2})$	$b_2[t_2^{A,1\prime}]$	=	$(B, t_1^{A,2\prime})$
	$b_2[t_2^{B,1}]$	=	$(C, t_1^{B,2})$	$b_2[t_2^{B,1\prime}]$	=	$(B, t_1^{B,2\prime})$
	$b_2[t_2^{C,1}]$	=	$(C, t_1^{C,2})$	$b_2[t_2^{C,1\prime}]$	=	$(B, t_1^{C,2\prime})$
	$b_2[t_2^D]$	=	$0.6(A, t_1^{A\prime}) + 0.4(B, t_1^{B\prime})$	$b_2[t_2^{D\prime}]$	=	$0.6(A, t_1^A) + 0.4(B, t_1^B)$
	$b_2[t_2^E]$	=	$(C, t_1^{C\prime})$	$b_2[t_2^{E\prime}]$	=	(C, t_1^C)

(c) Epistemic model scenario (*ii*)

Figure 8: Illustration proof Scenario (*ii*)

the belief $b_1^a = (c^1, d^7)$. The superscripts here in the belief indicate the order of belief the choice corresponds to.

Table 7: *Illustrating game for proof scenario (iii)*

Player 1's combinations of extreme first-order and seventh-order expectations

	(c,c)	(c,d)	(c,e)	(c,f)	(d,c)	(d,d)	...
a	0	1	0	0	0	0	0
b	1	1	1	1	1	1	1

Player 1's utilities

Player 2's extreme fourth-order expectations

	c	d	e	f
c	2	2	0	0
d	0	2	2	0
e	3	0	3	3
f	0	3	0	2

Player 2's utilities

We will focus in this scenario on making choice a optimal for a belief hierarchy expressing k-fold belief in rationality for every order of belief k on a path of player 1's causality diagram. To this end, first fix a type t_1^a for player 1.

Scenario (iii) differs from the previous ones in that we have multiple paths on the causality diagram of player 1. As a result, we cannot use the construction with sequences of probability one beliefs in exactly the same way as we did for scenarios (i) and (ii). To compare more specifically with the previous two scenarios, consider the following example. First, define type t_1^a to be such that it assigns probability one to the choice-type combination $(c, t_2^{a,1})$ for player 2. Let type $t_2^{a,1}$ on its turn assigns probability one to the choice-type combination $(\cdot, t_1^{a,2})$. Let type $t_1^{a,2}$ assign probability one to $(\cdot, t_2^{a,3})$; let type $t_2^{a,3}$ assign probability one to $(\cdot, t_1^{a,4})$; let type $t_1^{a,4}$ assign probability one to $(\cdot, t_2^{a,5})$; let type $(\cdot, t_2^{a,5})$ assign probability one to $(\cdot, t_1^{a,6})$; and let type $t_1^{a,6}$ assign probability one to the choice-type combination (d, t_2^d). Then clearly choice a is optimal given type t_1^a. However, player 1 would not express 1-fold belief in rationality and thus also not on-path belief in rationality. Namely, choice c is only optimal for a probabilistic fourth-order expectation. We have that t_1^a assigns probability one to the choice-type combination $(c, t_2^{a,1})$, from which there follows a sequence of three more probability one beliefs until type $t_1^{a,4}$, which induces a belief that assigns probability one to $(\cdot, t_2^{a,5})$. Whatever choice we fill in to complete the belief $b[t_1^{a,4}]$, choice c assigned probability one to in $b[t_1^a]$ will then never be optimal, as by construction the fourth-order expectation induced by $t_2^{a,1}$ will be non-probabilistic. Another route we could take is to make the belief $b[t_1^{a,4}]$ probabilistic, in such a way that choice c in the support of $b[t_1^a]$ becomes optimal. However, then we are not taking into account that the seventh-order expectation induced by type t_1^a should be non-probabilistic. That is, choice a is only optimal for the non-probabilistic belief (or seventh-order expectation) $b_1^a = (c^1, d^7)$. So we cannot fix just any type $t_1^{a,4}$ such that choice c is optimal given type $t_1^{a,1}$.

When constructing types for a partial epistemic model where type t_1^a expresses k-fold belief in rationality for every order k on the causality diagram, we therefore should at all times look at *combinations* of choices. As an (incomplete) illustration of such a partial epistemic model, consider Figure 9. In this figure, we have type t_1^a for player 1. All the remaining types have combinations of two choices in their superscripts. The reason for this is that the sequences of types we will construct in this step will be such that they optimize a combination of two choices. These choices appear in different orders of belief. For instance, we define type t_1^a now such that $b[t_1^a](c, t_2^{cf}) = 1$. The idea is to construct t_2^{cf} such that choice c for player 2 is optimal given this type. Then type t_1^a would express

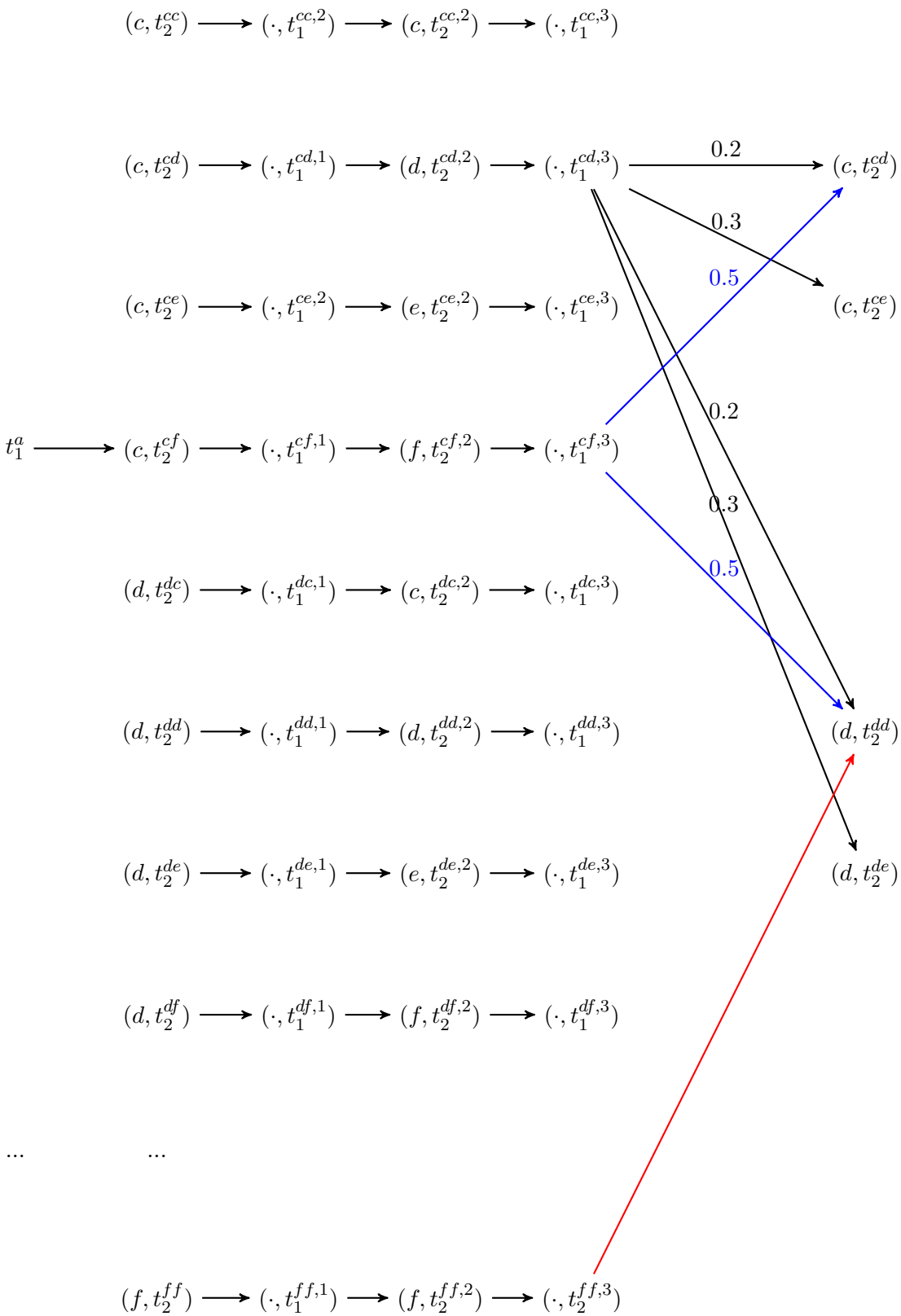


Figure 9: Step 1 scenario (*iii*)

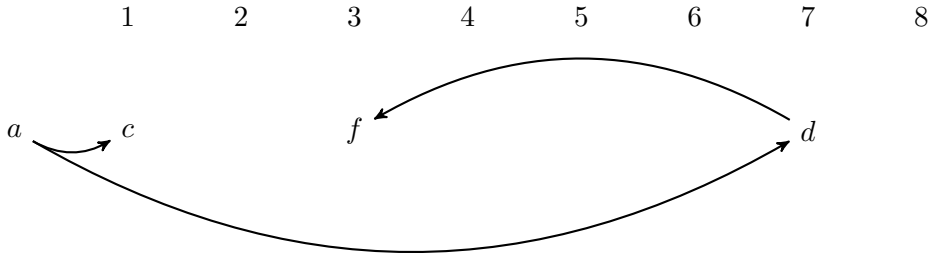


Figure 10: Step 1 scenario (iii)

1-fold belief in rationality. We do so by first constructing a sequence of types $(t_2^{cf}, t_1^{cf,1}, t_2^{cf,2}, t_1^{cf,3})$. In this sequence, we have $b_2[t_2^{cf}](\cdot, t_1^{cf,1}) = 1$, $b_1[t_1^{cf,1}](f, t_2^{cf,2}) = 1$ and $b_2[t_2^{cf,2}](\cdot, t_1^{cf,3}) = 1$. Finally, we have $b_1[t_1^{cf,3}] = 0.5(c, t_2^{cd}) + 0.5(d, t_2^{dd})$. The fourth-order expectation that type t_2^{cf} induces is then $b_2^c = 0.5c + 0.5d$. Looking at the utilities for player 2 in Table 7, this indeed makes choice c for player 2 optimal.

The second choice listed in the superscript of type t_2^{cf} is the choice f. When player 1 is of type t_1^a, her third-order expectation corresponds to a belief that assigns probability one to choice f. That is, we have that $b[t_1^{cf,1}](f, t_2^{cf,2}) = 1$. The reason why we look specifically at the third-order expectation here is illustrated in Figure 10. We know that choice a is only optimal given a seventh-order expectation that places probability one on d^7. Directly specifying the seventh-order expectation is problematic since the fifth-order expectation necessarily needs to be probabilistic. Therefore, what we do is the following. For each order of belief in N_1 beyond order $1+4 = 5$, we first reason backwards, up until we get in the range of beliefs between orders 1 and 5. In this case, this means we first reason backwards one optimality-relevant step from order 7 to order $7 - 4 = 3$. We fix one choice for player 2 that is optimal given a fourth-order expectation that places probability one on choice d^7. In this game the only optimal choice for such a fourth-order expectation is f^3. Hence we fix choice f. Note that if we had order 11 instead of 7, we would have to reason two steps backwards first: first from order 11 to order $11 - 4 = 7$ and then from order 7 to order $7 - 4 = 3$.

As we can conclude from Figure 9, choice f is optimal for type $t_2^{cf,2}$. That is, we have that $b_2[t_2^{cf,2}](\cdot, t_1^{cf,3}) = 1$ and that $b_1[t_1^{cf,3}] = 0.5(c, t_2^{cd}) + 0.5(d, t_2^{dd})$. In this model, we defined t_2^{cd} such that it induces a second-order expectation that places probability one on choice d. The same applies to type t_2^{dd}. Hence, the fourth-order expectation induced by type $t_2^{cf,2}$ is $b_2^f = 0.5d + 0.5d = d$.

We can construct such types like t_2^{cf} for each combination of choices in $C_2^\infty \times C_2^\infty$. To take another example of a combination of choices, consider (c^1, d^3). Here we define the type t_2^{cd} and the types in the sequence $(t_1^{cd,1}, t_2^{cd,2}, t_1^{cd,3})$ in a similar way as before. This is again illustrated in Figure 9. We have that choice c is optimal given a fourth-order expectation $b_2^c = 0.5c + 0.5d$. We have that choice d is optimal given a fourth-order expectation $b_2^d = 0.4d + 0.6e$. The resulting joint probability distribution is $b_2^{cd} = 0.5 \cdot 0.4 \cdot (c, d) + 0.5 \cdot 0.6 \cdot (c, e) + 0.5 \cdot 0.4 \cdot (d, d) + 0.5 \cdot 0.6 \cdot (d, e) = 0.2(c, d) + 0.3(c, e) + 0.2(d, d) + 0.3(d, e)$. Thus we define type $t_1^{cd,3}$ to be such that $b_1[t_1^{cd,3}] = 0.2(c, t_2^{cd}) + 0.3(c, t_2^{ce}) + 0.2(d, t_2^{d,d}) + 0.3(d, t_2^{d,e})$. By construction, type t_2^{cd} then induces a fourth-order expectation b_2^c and type $t_2^{cd,2}$ induces a fourth-order expectation b_2^d. Therefore choice c^1 is optimal given the type t_2^{cd} and choice d^3 is optimal given the type $t_2^{cd,2}$.

For each combination of two choices $(c_2^1, c_2^3) \in C_2^\infty \times C_2^\infty$, we now do the following. Take any such combination. Create a type for this combination: $t_2^{c_2^1 c_2^3}$. Define this type to be such that $b_2[t_2^{c_2^1 c_2^3}]$

assigns probability one to type $t_1^{c_2^1 c_2^3,1}$. Define type $t_1^{c_2^1 c_2^3,1}$ to be such that the belief $b_1[t_1^{c_2^1 c_2^3,1}]$ assigns probability one to $(c_2^3, t_2^{c_2^1 c_2^3,2})$. Define type $t_2^{c_2^1 c_2^3,2}$ such that $b_2[t_2^{c_2^1 c_2^3,2}]$ assigns probability one to $t_1^{c_2^1 c_2^3,3}$. Finally, for type $t_1^{c_2^1 c_2^3,3}$ we need to do the following. For every $c_2 \in C_2$, let $b_2^{c_2} \in \Delta(C_2^\infty)$ be a fourth-order expectation that makes c_2 optimal. Let $b_2^{c_2^1 c_2^3} \in \Delta(C_2^\infty \times C_2^\infty)$ be the product of $b_2^{c_2^1}$ and $b_2^{c_2^3}$. Now, we define $b_1[t_1^{c_2^1 c_2^3}(3)]$ in the following manner:

$$b_1[t_1^{c_2^1 c_2^3,3}](c_2^{1\prime}, t_2^{c_2^{1\prime} c_2^{3\prime}}) = b_2^{c_2^1 c_2^3}(c_2^{1\prime}, c_2^{3\prime}), \tag{1}$$

for each combination of choices $(c_2^{1\prime}, c_2^{3\prime}) \in C_2^\infty \times C_2^\infty$.

Construct such sequences of types for every possible combination of player 2's choices in the product-space $C_2^\infty \times C_2^\infty$. This results in a partial epistemic model as is partially illustrated in Figure 9. First, each type $t_2^{c_2^1 c_2^3}$ in this model is such that that choice c_2^1 is optimal given this type. Namely, the fourth-order expectation of type $t_2^{c_2^1 c_2^3}$ is $b_2^{c_2^1}$ by equation (1). Also, choice c_2^3 is optimal for type $t_2^{c_2^1 c_2^3,2}$, because the fourth-order expectation of type $t_2^{c_2^1 c_2^3,2}$ is $b_2^{c_2^3}$ by equation (1). Second, each type $t_2^{c_2^1 c_2^3}$ in this model expresses 4-fold belief in rationality. Namely, the final type in the sequence of types $t_1^{c_2^1 c_2^3,3}$ is such that it only assigns positive probability to choice-type combinations $(c_2^{1\prime}, t_2^{c_2^{1\prime} c_2^{3\prime}})$. By construction, we have in such choice-type combinations that the choice $c_2^{1\prime}$ is optimal given the type $t_2^{c_2^{1\prime} c_2^{3\prime}}$. Similarly, we also have by construction that each type $t_2^{c_2^1 c_2^3,2}$ expresses 4-fold belief in rationality: from each such type, "following" four arrows in Figure 9 always brings us to choice-type combinations $(c_2^{3\prime}, t_2^{c_2^{1\prime} c_2^{3\prime},2})$. As shown before, we have in such choice-type combination that the choice $c_2^{3\prime}$ is optimal given the type $t_2^{c_2^{1\prime} c_2^{3\prime},2}$. Because the types $t_2^{c_2^1 c_2^3}$ and $t_2^{c_2^1 c_2^3,2}$ for each combination of choices (c_2^1, c_2^3) expresses 4-fold belief in rationality and because $N_2 = \{4\}$, we have in fact that each such type for player 2 also expresses k-fold belief in rationality for every k on a path in player 2's causality diagram.

Now, recall that $N_1 = \{1, 7\}$. Looking at Figure 9, one arrow away from type t_1^a we have the choice-type combination (c, t_2^{cf}) and seven arrows away from type t_1^a we have the choice-type combinations $(d, t_2^{cd,2})$ and $(d, t_2^{dd,2})$. Hence, the first-order and seventh-order expectation induced by type t_1^a are (c^1, d^7), which means that a is optimal for t_1^a. Also, type t_1^a expresses 1-fold and 7-fold belief in rationality and believes that player 2 expresses k-fold belief in rationality for every k on a path in her causality diagram. It then follows that also player 1 expresses k-fold belief in rationality for every order k on a path in her causality diagram if she is of type t_1^a.

The model illustrated in Figure 9 is not completed. We only completed the sequences of types for the following combinations of choices: (c, d), (c, f) and (f, f). We leave the remainder of the illustration of Step 1 to the reader. The backward construction of Step 2 is almost completely analogous to Step 2 for scenarios (i) and (ii). For more details the reader is referred to the Appendix.

From Lemma 4 we conclude that if the causality diagram of player 1 is overlap-free, then each choice that survives the IESDC-procedure for player 1 is also a rational choice under common belief in rationality. From Lemma 3 we concluded the reverse. Together then, Lemma 3 and Lemma 4 prove Theorem 2: the IESDC-procedure exactly characterizes the rational choices under common belief in rationality for player 1 if and only if the player 1's causality diagram is overlap-free. We

know the causality diagram is overlap-free if and only if at least one of the cases as is listed in Theorem 2 is true.

7 Concluding remarks

Since its introduction by Geanakoplos et al. (1989), psychological game theory has proven to be a competent framework to model many belief-dependent motivations in games. Much of the work in this framework illustrates that, compared to traditional games, reasoning about and in situations with belief-dependent motivations can be rather complex. Some properties of traditional games that add intuition to reasoning in such games do not always carry over to psychological games. In this paper we focused on one of such failures: the exact characterization of common belief in rationality by the iterated elimination of strictly dominated choices (IESDC) procedure. The IESDC procedure has proven to be a very useful algorithm to analyse traditional games as it is straightforward to use and an intuitive notion because of its characterization of common belief in rationality. This sparked the question in what kind of psychological games the IESDC-procedure always characterizes rational choices under common belief in rationality. By exactly identifying these cases, we also wished to give intuition as to why the IESDC-procedure may fail in other cases.

The IESDC-procedure takes into account that players may have belief-dependent motivations. The manner in which decision problems are defined clearly lets utilities depend on higher-order expectations. In each elimination round, for those choices that survive the round we can always find *some* belief hierarchy in the (reduced) decision problem such that the relevant choice is optimal. We do this for each (reduced) decision problem independently. The complexity psychological games introduce is that reasoning steps may overlap, as we illustrated via causality diagrams. In order for a belief hierarchy to express k-fold belief in rationality, restrictions need to be imposed on particular higher-order beliefs. Expressing k'-fold belief in rationality may require restrictions on the same higher-order belief. These restrictions can be in conflict. Thus even though a choice (1) may be rational under a belief hierarchy expressing k-fold belief in rationality and (2) that same choice may be rational under a belief hierarchy expressing k'-fold belief in rationality, this choice may not be rational under any belief hierarchy that expresses both k-fold and k'-fold belief in rationality. The IESDC-procedure cannot take into account this friction. To the extent the IESDC-procedure characterizes particular reasoning steps of an individual, it does so for each such reasoning step independently. This completely disregards any overlap in reasoning steps.

When a causality diagram for a player is overlap-free, contradicting restrictions on the same order of belief in order to express common belief in rationality cannot occur. The main result that we have shown in this paper is that in precisely such cases IESDC always characterizes the rational choices under common belief in rationality.

In total we identified three cases in which causality diagrams are overlap-free. These include two relatively trivial cases in the sense that the causality diagram has a single path, and one non-trivial case. Though interesting kinds of psychological games can be captured by these three, it can be argued that many types of psychological games that are prominent in practice cannot. Namely, we have that if both players in an expectation-based psychological game care for the material outcome of the game and at least one player has some belief-dependent motivation, already then the IESDC-procedure is not guaranteed to characterize the rational choices under common belief in rationality. In particular in experimental settings this will very often be the case, as subjects need to be incentivized by material pay-offs.

A couple of natural extensions of the questions asked in this paper arise. First, to keep matters tractable we focused on two-player expectation-based psychological games. We did not venture into the topic of expectation-based psychological games with many players. One interesting additional complexity such settings bring along is the question of whether correlation between the beliefs of opponents' matters in the formation of higher-order expectations.

A second natural extension of this research would be to consider similar questions as asked in this paper for dynamic psychological games. Many instances of belief-dependent motivations arise when players have the opportunity to learn about the beliefs and intentions of their opponents by observing their past behaviour. Such instances can also arise in one-stage games, where the updated belief after play can be utility-relevant as well (Battigalli and Dufwenberg, 2009). In traditional settings, prominent reasoning concepts for dynamic games are for instance common belief in future rationality (which can capture backward induction reasoning) as in Perea (2014) and common strong belief in rationality (which captures forward induction reasoning) as in Battigalli and Siniscalchi (2002). Both concepts are characterized by procedures that (amongst other things) rely on iteratively eliminating strictly dominated strategies. The natural question then arises to what extent such elimination procedures also succeed in characterizing relevant reasoning concepts in expectation-based psychological games.

Throughout this paper we assumed common knowledge of players' motivations, including belief-dependent motivations. It is a strong assumption to make that psychological entities such as belief-dependent motivations are completely transparant among all players in a game (Attanasi et al., 2016). Elimination procedures have already been developed for traditional games with incomplete information (see for instance Bach and Perea (2016)). A final extension one therefore could consider is how well such elimination procedures fare in characterizing the relevant rationality concepts in expectation-based psychological games with incomplete information.

Appendix

A Proof Lemma 4

We recall here Lemma 4.

Lemma 4. *Consider a family of games $\mathcal{G}(N_1, N_2)$. If the causality diagram of player 1 is overlap-free, then for every game in $\mathcal{G}(N_1, N_2)$, each choice that survives the IESDC-procedure for player 1 is also a rational choice under common belief in rationality.*

Proof. We will now prove this theorem for the three scenarios described in the main text. We will do so in two steps. We will take some choice $c_1 \in C_1^\infty$ that survives the IESDC-procedure. Then in Step 1 we will create a partial epistemic model with a type that makes choice c_1 optimal. Moreover, we construct this type such that for each order k on a path in player 1's causality diagram, this type expresses k-fold belief in rationality. Afterwards, in Step 2, we will show that from *any* partial epistemic model including a type as created in Step 1, we can create a full epistemic model with a type that makes choice c_1 optimal and that expresses common belief rationality. We do so by making use of a backward, recursive procedure that in each iteration simultaneously constructs types and choices (which are optimal given those types).

Scenario (i)
Step 1
First consider scenario (i) with $N_1 = \{a\}$ and a even. In this Step 1, we will construct a partial epistemic model. By *partial* we mean we only completely specify the beliefs induced for some particular types.

For each choice $c_1 \in C_1^\infty$, fix an a-th order expectation $b_1^{c_1} \in \Delta(C_1^\infty)$ for which c_1 is optimal.[4] The reason we can do so is as follows. From Lemma 1 we know that for each choice that is not strictly dominated in a decision problem, we can find a belief in that decision problem such that the relevant choice is optimal. The final reduced decision problem resulting from the IESDC-procedure leaves the choices in C_1^∞ for player 1 in the decision problem. That is, because order a is even we have as a reduced decision problem after following through with the IESDC-procedure: $(C_1^\infty, C_1^\infty, v_1)$. By Lemma 1, we should then have that each choice in C_1^∞ is optimal for some belief in $\Delta(C_1^\infty)$.

Subsequently, for each $c_1 \in C_1^\infty$, construct a type $t_1^{c_1}[c_1]$. Take for each $t_1^{c_1}[c_1]$ a sequence of types $(t_1^{c_1}[c_1], t_2^{c_1,1}[c_1], ..., t_2^{c_1,a-1}[c_1])$. Then, for each $c_1 \in C_1^\infty$ let us have in each such sequence that type $t_1^{c_1}[c_1]$ assigns probability one to type $t_2^{c_1,1}[c_1]$, and that type $t_i^{c_1,n}[c_1]$ with $i \in \{1, 2\}$ assigns probability one to type $t_j^{c_1,n+1}$ with $j \neq i$, for each $n \in \{1, ..., a-2\}$.

Next, for each $c_1 \in C_1^\infty$ construct for each $c_1' \in C_1 \infty$ a type $t_1^{c_1,a}[c_1']$. Then define $t_2^{c_1,a-1}[c_1]$ to be such that

$$b_2[t_2^{c_1,a-1}[c_1]](c_1', t_1^{c_1,a}[c_1'']) := \begin{cases} b_1^{c_1}(c_1'), \text{ if } c_1' = c_1'' \\ 0, \text{ otherwise.} \end{cases}$$

Now, we do a similar thing $k^* - 2$ times, where $k^* = max(N_1 \cup N_2)$. For each $p \in \{1, ..., k^* - 2\}$, do

[4] With a-th order expectation in this context we specifically refer to $marg_{C_1} e_1^a \in \Delta(C_1)$ where $e_1^a \in \Delta(W_1^{a-1} \times C_1)$.

the following: For each $c_1, c_1' \in C_1^\infty$ take a sequence of types $(t_1^{c_1,pa}[c_1'], t_2^{c_1,pa+1}[c_1'], ..., t_2^{c_1,(p+1)a-1}[c_1'])$. Then, let us have in each such sequence that type $t_1^{c_1,pa}[c_1']$ assigns probability one to type $t_2^{c_1,pa+1}[c_1']$, and that type $t_i^{c_1,pa+n}[c_1']$ with $i \in \{1,2\}$ assigns probability one to type $t_j^{c_1,pa+n+1}$ with $j \neq i$, for each $n \in \{1, ..., a-2\}$.

Then, for each $c_1, c_1' \in C_1^\infty$ construct a type $t_1^{c_1,(p+1)a}[c_1']$. Then define $t_2^{c_1,(p+1)a-1}[c_1']$ to be such that

$$b_2[t_2^{c_1,(p+1)a-1}[c_1']](\bar{c}_1, t_1^{c_1,(p+1)a}[c_1'']) := \begin{cases} b_1^{c_1'}(\bar{c}_1), \text{ if } \bar{c}_1 = c_1'' \\ 0, \text{ otherwise.} \end{cases}$$

Finally, consider the case $p = k^* - 1$. For each combination of choices $c_1, c_1' \in C_1^\infty$ take a sequence of types $(t_1^{c_1,(k^*-1)a}[c_1'], t_2^{c_1,(k^*-1)a+1}[c_1'], ..., t_2^{c_1,k^*a-1}[c_1'])$. Then, let us have in each such sequence that type $t_1^{c_1,(k^*-1)a}[c_1']$ assigns probability one to type $t_2^{c_1,(k^*-1)a+1}[c_1']$, and that type $t_i^{c_1,(k^*-1)a+n}[c_1']$ with $i \in \{1,2\}$ assigns probability one to type $t_j^{c_1,(k^*-1)a+n+1}$ with $j \neq i$, for each $n \in \{1, ..., a-2\}$.

Finally define $t_2^{c_1,k^*a-1}[c_1']$ to be such that

$$b_2[t_2^{c_1,k^*a-1}[c_1']](\bar{c}_1, t_1^{c_1''}[c_1'']) := \begin{cases} b_1^{c_1'}(\bar{c}_1), \text{ if } \bar{c}_1 = c_1'' \\ 0, \text{ otherwise.} \end{cases}$$

So we have that the distribution over choices induced by type $t_2^{c_1,a-1}[c_1]$ is equal to the distribution over choices represented by the a-th order expectation $b_1^{c_1}$. From type $t_1^{c_1}[c_1]$ there follows a sequence of probability one beliefs up to type $t_2^{c_1,a-1}[c_1]$. The a-th order expectation induced by type $t_1^{c_1}[c_1]$ is thus equal to $b_1^{c_1}$. Taken together, then choice c_1 is optimal given type $t_1^{c_1}$.

A similar line of reasoning holds for each choice c_1' in combination with the type $t_1^{c_1,pa}[c_1']$, for each $c_1 \in C_1^\infty$ and each $p \in \{1, ..., k^* - 1\}$. The distribution over choices induced by type $t_2^{c_1,(p+1)a-1}[c_1']$ is equal to the distribution over choices represented by the a-th order expectation $b_1^{c_1'}$. From type $t_1^{c_1,pa}[c_1]'$ there follows a sequence of probability one beliefs up to type $t_2^{c_1,(p+1)a-1}[c_1']$. The a-th order expectation induced by type $t_1^{c_1,pa}[c_1']$ is thus equal to $b_1^{c_1'}$. Taken together, then choice c_1' is optimal given type $t_1^{c_1,pa}[c_1']$.

We do the above for each $c_1 \in C_1^\infty$. Call the resulting partial epistemic model $\mathcal{M}$. By construction, we have for each $c_1 \in C_1^\infty$ that c_1 is optimal given $t_1^{c_1}[c_1]$. Moreover, each $t_1^{c_1}[c_1]$ also expresses on-path belief in rationality. This is because for each order of belief pa for $p \in \{1, ..., k^* - 1\}$ the type $t_2^{c_1,pa-1}[c_1']$ (with $t_2^{c_1,a-1}[c_1]$ for $p = 1$ specifically) only assigns positive probability to choice type pairs $(c_1'', t_1^{c_1,(p+1)a}[c_1''])$. In these pairs the choice is optimal for the type by construction. Additionally, for order k^*a, type $t_2^{c_1,k^*a-1}[c_1']$ only assigns positive probability to choice type pairs $(c_1'', t_1^{c_1''}[c_1''])$. Also in these pairs the choice is optimal for the type by construction.

Step 2

In Step 1 we have shown that for every choice $c_1 \in C_1^\infty$ we can always construct a *partial* epistemic model with a type $t_1^{c_1}[c_1]$ for which c_1 is optimal and that expresses on-path belief in rationality. In Step 2 we will now do the following. We will show that if there exists a belief hierarchy expressing on-path belief in rationality for which c_1 is optimal, then there is also a belief hierarchy expressing common belief in rationality for which c_1 is optimal.

Consider a partial epistemic model $\mathcal{M} = (T_i, b_i[t_i])_{i \in \{1,2\}}$ as constructed in Step 1 with a type $t_1^{c_1}[c_1]$ that expresses on-path belief in rationality and for which c_1 is optimal. By means of a backward, recursive procedure we transform this epistemic model such that we get to a new, complete

epistemic model that includes a type $t_1^m[c_1, c_1, 0]$ that expresses common belief in rationality and induces the same a-th order expectation as type $t_1^{c_1}[c_1]$ does. The recursive procedure here defines choices *and* types at the same time in each iteration.

The recursive procedure is as follows.

Iteration 0: For each choice $c_1 \in C_1^\infty$, define

$$d^0[c_1, c_1, 0] := c_1.$$

Moreover, for each choice $c_1 \in C_1^\infty$ and each $k \in \{1, 2, ..., a-1\}$, define $d^0[c_1, c_1, k]$ randomly. So

$$d^0[c_1, c_1, k] := c', \text{ for } \textit{some } c' \in C_1^\infty \text{ if } k \text{ is even or } \textit{some } c' \in C_2^\infty \text{ if } k \text{ is odd.}$$

For each $p \in \{1, ..., k^* - 1\}$ let us have in a similar fashion that

$$d^0[c_1, c_1', pa] := c_1',$$

and for every $k \in \{1, 2, ..., a-1\}$ that

$$d^0[c_1, c_1', pa + k] := c', \text{ for } \textit{some } c' \in C_1^\infty \text{ if } k \text{ is even or } \textit{some } c' \in C_2^\infty \text{ if } k \text{ is odd.}$$

Take a sequence of types $(t_1^0[c_1, c_1, 0], t_2^0[c_1, c_1, 1], ..., t_2^0[c_1, c_1, a-1])$ for every choice $c_1 \in C_1^\infty$. Similarly, for each $p \in \{1, ..., k^* - 1\}$ and each pair of choices $c_1, c_1' \in C_1^\infty$, take a sequence of types $(t_1^0[c_1, c_1', pa], t_2^0[c_1, c_1', pa + 1], ..., t_2^0[c_1, c_1', (p+1)a - 1])$.

Now, for each $c_1 \in C_1^\infty$, define type $t_1^0[c_1, c_1, 0]$ such that

$$b_1[t_1^0[c_1, c_1, 0]] := (d^0[c_1, c_1, 1], t_2^0[c_1, c_1, 1]).$$

Then, define for each $k \in \{1, 2, ..., a-2\}$ type $t_i^0[c_1, c_1, k]$ with $i \in \{1, 2\}$ to be such that

$$b_i[t_i^0[c_1, c_1, k]] := (d^0[c_1, c_1, k+1], t_j^0[c_1, c_1, k+1]).$$

Finally, we define for each choice $c_1 \in C_1^\infty$ type $t_2^0[c_1, c_1, a-1]$ to be such that

$$b_2[t_2^0[c_1, c_1, a-1]](c_1', t_1^0[c_1, c_1', 0]) := b_1^{c_1}(c_1'), \forall c_1' \in C_1^\infty.$$

Similarly, for each $c_1, c_1' \in C_1^\infty$ and each $p \in \{1, ..., k^* - 1\}$ define $t_1^0[c_1, c_1', pa]$ to be such that

$$b_1[t_1^0[c_1, c_1', pa]] := (d^0[c_1, c_1', pa + 1], t_2^0[c_1, c_1', pa + 1]).$$

And define for each $k \in \{1, 2, ..., a-2\}$ type $t_i^0[c_1, c_1', pa + k]$ to be such that

$$b_i[t_i^0[c_1, c_1', pa + k]] := (d^0[c_1, c_1', pa + k + 1], t_j^0[c_1, c_1', pa + k + 1]).$$

Finally, we define type $t_2^0[c_1, c_1', (p+1)a - 1]$ for $p \in \{1, ..., k^* - 2\}$ to be such that

$$b_2[t_2^0[c_1, c_1', (p+1)a - 1]](c_1'', t_1^0[c_1, c_1'', (p+1)a]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

If $p = k^* - 1$, define type $t_2^0[c_1, c_1', k^*a - 1]$ to be such that

$$b_2[t_2^0[c_1, c_1', k^*a - 1]](c_1'', t_1^0[c_1'', c_1'', 0]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

Note that by construction of Step 1, we have for each $p \in \{1, ..., k^*\}$ that $b_1^{c_1'}(c_1'') = b_2[t_2^0[c_1, c_1', pa-1]](c_1'', t_1^0[c_1, c_1'', pa])$ for each $c_1'' \in C_1^\infty$. Moreover, all other types induce a probability one belief. This implies that type $t_1^0[c_1, c_1, 0]$ induces exactly the same a-th order expectation as type $t_1^{c_1}[c_1]$ did in Step 1. Similarly, each type $t_1^0[c_1, c_1', pa]$ induces the same a-th order expectation as type $t_1^{c_1, pa}[c_1']$ did. So for Iteration 0 we essentially take a copy of the epistemic model created in Step 1, but fill in the beliefs that were still incomplete from this step.

Iteration n $\geqslant$ 1 : For each choice $c_1 \in C_1^\infty$ and each choice $c_1' \in C_1^\infty$ define type $t_2^n[c_1, c_1', k^*a - 1]$ to be such that

$$b_2[t_2^n[c_1, c_1', k^*a - 1]](c_1'', t_1^{n-1}[c_1'', c_1'', 0]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

For each $c_1, c_1' \in C_1^\infty$, we then also define

$$d^n[c_1, c_1', k^*a - 1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_1, c_1', k^*a - 1].$$

Now, for each pair of choices $c_1, c_1' \in C_1^\infty$, define recursively for each *even* $k \in \{2, ..., a-2\}$ starting at $k = a - 2$, type $t_1^n[c_1, c_1', (k^* - 1)a + k]$ to be such that

$$b_1[t_1^n[c_1, c_1', (k^* - 1)a + k]] := (d^n[c_1, c_1', (k^* - 1)a + k + 1], t_2^n[c_1, c_1', (k^* - 1)a + k + 1]).$$

Second, also define

$$d^n[c_1, c_1', (k^* - 1)a + k] := \bar{c}_1, \text{ with } \bar{c}_1 \text{ optimal given the type } t_1^n[c_1, c_1', (k^* - 1)a + k].$$

Third, define type $t_2^n[c_1, c_1', (k^* - 1)a + k - 1]$ to be such that

$$b_2[t_2^n[c_1, c_1', (k^* - 1)a + k - 1]] := (d^n[c_1, c_1', (k^* - 1)a + k], t_1^n[c_1, c_1', (k^* - 1)a + k]).$$

Fourth, also define

$$d^n[c_1, c_1', (k^* - 1)a + k - 1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_1, c_1', (k^* - 1)a + k - 1].$$

Finally, for each $c_1, c_1' \in C_1^\infty$ define type $t_1^n[c_1, c_1', (k^* - 1)a]$ to be such that

$$b_1[t_1^n[c_1, c_1', (k^* - 1)a]] := (d^n[c_1, c_1', (k^* - 1)a + 1], t_2^n[c_1, c_1', (k^* - 1)a + 1]),$$

and define

$$d^n[c_1, c_1', (k^* - 1)a], (k^* - 1)a] := c_1'.$$

Next, for each $p \in \{0, ..., k^* - 2\}$, do the following iteratively, going backwards starting at $p = k^* - 2$:

For each choice $c_1 \in C_1^\infty$ and each choice $c_1' \in C_1^\infty$ define type $t_2^n[c_1, c_1', (p+1)a - 1]$ to be such that

$$b_2[t_2^n[c_1, c_1', (p+1)a - 1]](c_1'', t_1^n[c_1, c_1'', (p+1)a]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

For each $c_1, c_1' \in C_1^\infty$, we then also define

$$d^n[c_1, c_1', (p+1)a - 1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_1, c_1', (p+1)a - 1].$$

Now, for each pair of choices $c_1, c_1' \in C_1^\infty$, define recursively for each *even* $k \in \{2, ..., a-2\}$ starting at $k = a - 2$, type $t_1^n[c_1, c_1', pa + k]$ to be such that

$$b_1[t_1^n[c_1, c_1', pa + k]] := (d^n[c_1, c_1', pa + k + 1], t_2^n[c_1, c_1', pa + k + 1]).$$

Second, also define

$$d^n[c_1, c_1', pa + k] := \bar{c}_1, \text{ with } \bar{c}_1 \text{ optimal given the type } t_1^n[c_1, c_1', pa + k].$$

Third, define type $t_2^n[c_1, c_1', pa + k - 1]$ to be such that

$$b_2[t_2^n[c_1, c_1', pa + k - 1]] := (d^n[c_1, c_1', pa + k], t_1^n[c_1, c_1', pa + k]).$$

Fourth, also define

$$d^n[c_1, c_1', pa + k - 1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_1, c_1', pa + k - 1].$$

Finally, for each $c_1, c_1' \in C_1^\infty$ define type $t_1^n[c_1, c_1', pa]$ to be such that

$$b_1[t_1^n[c_1, c_1', pa]] := (d^n[c_1, c_1', pa + 1], t_2^n[c_1, c_1', pa + 1]),$$

and define

$$d^n[c_1, c_1', pa] := c_1'.$$

We do this iteratively for each $p \in \{0, ..., k^* - 2\}$, starting at $p = k^* - 2$.

We have that C_1^∞ and C_2^∞ are finite sets. Moreover, a and k^* are finite orders of belief, and therefore k^*a is as well. Hence, there are iterations m, n with $m > n$ such that:

$$d^m[c_1, c_1, k] = d^n[c_1, c_1, k], \ \forall c_1 \in C_1^\infty, k \in \{0, 1, ..., a - 1\},$$

and

$$d^m[c_1, c_1', pa + k] = d^n[c_1, c_1', pa + k], \ \forall c_1, c_1' \in C_1^\infty, k \in \{0, 1, ..., a - 1\}, p \in \{1, ..., k^* - 1\}.$$

When we find such iterations m and n, we stop the recursive procedure.

Now we create the epistemic model $\mathcal{M}^*$ from the types we have constructed in our recursive procedure. Define $T_1(l) := \{t_1^l[c_1, c_1, k] : c_1 \in C_1^\infty, k \in \{0, ..., a - 2\} \text{ even}\} \cup \{t_1^l[c_1, c_1', pa + k] : c_1, c_1' \in C_1^\infty, \ p \in \{1, ..., k^* - 1\}, \ k \in \{0, ..., a - 2\} \text{ even}\}$ and $T_2(l) := \{t_2^l[c_1, c_1, k] : c_1 \in C_1^\infty, k \in \{1, ..., a - 1\} \text{ odd}\} \cup \{t_2^l[c_1, c_1', pa + k] : c_1, c_1' \in C_1^\infty, \ p \in \{1, ..., k^* - 1\}, \ k \in \{1, ..., a - 1\} \text{ odd}\}$. Then, let $T(l) := T_1(l) \cup T_2(l)$. Do this for every $l \in \{n, ..., m\}$.

In $T(n + 1)$ specifically, we re-define for each $c_1, c_1' \in C_1^\infty$ the type $t_2^{n+1}[c_1, c_1', k^*a - 1]$ to be such that

$$b_2[t_2^{n+1}[c_1, c_1', k^*a - 1]](c_1'', t_1^m[c_1'', c_1'', 0]) := b_1^{c_1'}(c_1''), \forall c_1'' \in C_1^\infty.$$

So instead of assigning positive probability to types in $T(n)$, each type $t_2^{n+1}[c_1, c_1', k^*a - 1]$ now assigns positive probability to types in $T(m)$. Then define $\mathcal{M}^* := (\bigcup_{l \in \{n+1, ..., m\}} T_i(l), b[t_i])_{i \in \{1,2\}}$.

We will show that each type in $\mathcal{M}^*$ expresses common belief in rationality. We will do so in steps.

First, we can note that for each $c_1 \in C_1^\infty$ and each $l \in \{n+1, ..., m\}$ in $\mathcal{M}^*$, choice c_1 is optimal for type $t_1^l[c_1, c_1, 0]$, and that for each $c_1, c_1' \in C_1^\infty$, each $l \in \{n+1, ..., m\}$ and each $p \in \{1, ..., k^* - 1\}$ choice c_1' is optimal for type $t_1^l[c_1, c_1', pa]$.

Namely, from type $t_1^l[c_1, c_1, 0]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_1^l[c_1, c_1, 0], t_2^l[c_1, c_1, 1], ..., t_1^l[c_1, c_1, a-2])$. This sequence of probability one beliefs ends at type $t_2^l[c_1, c_1, a-1]$. By construction, we have that

$$marg_{C_1^\infty} b_2[t_2^l[c_1, c_1, a-1]] = b_1^{c_1}.$$

It follows then that type $t_1^l[c_1, c_1, 0]$ induces an a-th order expectation that is equal to $b_1^{c_1}$. We constructed $b_1^{c_1}$ such that c_1 is optimal given $b_1^{c_1}$. Hence c_1 is optimal given type $t_1^l[c_1, c_1, 0]$. This goes for every $l \in \{n+1, ..., m\}$.

Similarly for each $p \in \{1, ..., k^* - 1\}$, from type $t_1^l[c_1, c_1', pa]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_1^l[c_1, c_1', pa], t_2^l[c_1, c_1', pa+1], ..., t_1^l[c_1, c_1', (p+1)a-2])$. This sequence of probability one beliefs ends at type $t_2^l[c_1, c_1', (p+1)a-1]$. By construction, we have that

$$marg_{C_1^\infty} b_2[t_2^l[c_1, c_1', (p+1)a-1]] = b_1^{c_1'}.$$

It follows then that type $t_1^l[c_1, c_1', pa]$ induces an a-th order expectation that is equal to $b_1^{c_1'}$. We constructed $b_1^{c_1'}$ such that c_1' is optimal given $b_1^{c_1'}$. Hence c_1' is optimal given type $t_1^l[c_1, c_1', pa]$. This goes for every $l \in \{n+1, ..., m\}$.

Second, we can also show the following is true.

Claim 2. *Consider the epistemic model $\mathcal{M}^*$. For each $l \in \{n+1, ..., m\}$, each $k \in \{1, 2, ..., a-1\}$ and each $c_1 \in C_1^\infty$, choice $d^l[c_1, c_1, k]$ is optimal given the type $t_i^l[c_1, c_1, k]$ with $i \in \{1, 2\}$. Moreover, for each $p \in \{1, ..., k^* - 1\}$, each $l \in \{n+1, ..., m\}$, each $k \in \{1, 2, ..., a-1\}$ and each $c_1, c_1' \in C_1^\infty$, choice $d^l[c_1, c_1', pa+k]$ is optimal given the type $t_i^l[c_1, c_1', pa+k]$ with $i \in \{1, 2\}$.*

Proof of claim. We start off with the epistemic model we created when ending the recursive procedure, but *before* $\mathcal{M}^*$ was created.

For each $k \in \{0, 1, ..., a-2\}$ and each $c_1' \in C_1^\infty$ we have by construction that

$$b_i[t_i^n[c_1', c_1', k]](d^n[c_1', c_1', k+1], t_j^n[c_1', c_1', k+1]) = 1 = b_i[t_i^m[c_1', c_1', k]](d^m[c_1', c_1', k+1], t_j^m[c_1', c_1', k+1]),$$

with $d^n[c_1', c_1', k+1] = d^m[c_1', c_1', k+1]$. Note that these were the n and m that determined when to stop our recursive procedure. Moreover, for each $k \in \{0, 1, ..., a-2\}$, each $c_1', \bar{c}_1 \in C_1^\infty$ and each $p \in \{1, ..., k^* - 1\}$ we also have by construction

$$b_i[t_i^n[c_1', \bar{c}_1, pa+k]](d^n[c_1', \bar{c}_1, pa+k+1], t_j^n[c_1', \bar{c}_1, pa+k+1]) = 1 =$$
$$b_i[t_i^m[c_1', \bar{c}_1, pa+k]](d^m[c_1', \bar{c}_1, pa+k+1], t_j^m[c_1', \bar{c}_1, pa+k+1]),$$

with $d^n[c_1', \bar{c}_1, pa+k+1] = d^m[c_1', \bar{c}_1, pa+k+1]$. Additionally, we have by construction that

$$b_2[t_2^n[c_1', c_1', a-1]](d^n[c_1', c_1'', a], t_1^n[c_1', c_1'', a]) = b_1^{c_1'}[c_1''] = b_2[t_2^m[c_1', c_1', a-1]](d^m[c_1', c_1'', a], t_1^m[c_1', c_1'', a]),$$

for each $c_1'' \in C_1^\infty$. For each $p \in \{1, ..., k^* - 2\}$ we also have that

$$b_2[t_2^n[c_1', \bar{c}_1, pa-1]](d^n[c_1', c_1'', pa], t_1^n[c_1', c_1'', pa]) = b_1^{\bar{c}_1}[c_1''] =$$
$$b_2[t_2^m[c_1', \bar{c}_1, pa-1]](d^m[c_1', c_1'', pa], t_1^m[c_1', c_1'', pa]),$$

for each $c_1'' \in C_1^\infty$. Finally, we have that

$$b_2[t_2^n[c_1', \bar{c}_1, k^*a - 1]](d^{n-1}[c_1'', c_1'', 0], t_1^{n-1}[c_1'', c_1'', 0]) = b_1^{\bar{c}_1}[c_1''] =$$
$$b_2[t_2^m[c_1', \bar{c}_1, k^*a - 1]](d^{m-1}[c_1'', c_1'', 0], t_1^{m-1}[c_1'', c_1'', 0]),$$

for each $c_1'' \in C_1^\infty$.

Then, for each $c_1' \in C_1^\infty$, the pair of types $t_1^m[c_1', c_1', 0]$ and $t_1^n[c_1', c_1', 0]$ induce the same k^*a-th order *belief*. To see why this is the case, we can employ a recursive argument, for each $p \in \{1, ..., k^* - 1\}$ starting at $p = k^* - 1$.

We can first note that the pair of types $t_1^m[c_1', \bar{c}_1, (k^* - 1)a]$ and $t_1^n[c_1', \bar{c}_1, (k^* - 1)a]$ for each $c_1', \bar{c}_1 \in C_1^\infty$ induce the same a-th order belief. Namely, from the beginning of the proof of this claim we know that types $t_i^m[c_1', \bar{c}_1, (k^* - 1)a + k]$ and $t_i^n[c_1', \bar{c}_1, (k^* - 1)a + k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., a - 2\}$ induce a probability one belief. Moreover, the first-order belief induced by type $t_i^m[c_1', \bar{c}_1, (k^* - 1)a + k]$ for each $k \in \{1, ..., a - 1\}$ is equal to the first-order belief induced by type $t_i^n[c_1', \bar{c}_1, (k^* - 1)a + k]$. As a result, types $t_1^m[c_1', \bar{c}_1, (k^* - 1)a]$ and $t_1^n[c_1', \bar{c}_1, (k^* - 1)a]$ induce the same a-th order *belief*.

Now recall, for each $c_1', \bar{c}_1 \in C_1^\infty$, we have that

$$b_2[t_2^n[c_1', \bar{c}_1, (k^* - 1)a - 1]](d^n[c_1', c_1'', (k^* - 1)a], t_1^n[c_1', c_1'', (k^* - 1)a]) = b_1^{\bar{c}_1}[c_1''] =$$
$$b_2[t_2^m[c_1', \bar{c}_1, (k^* - 1)a - 1]](d^m[c_1', c_1'', (k^* - 1)a], t_1^m[c_1', c_1'', (k^* - 1)a])$$

Both types $t_2^m[c_1', \bar{c}_1, (k^*-1)a-1]$ and $t_2^n[c_1', \bar{c}_1, (k^*-1)a-1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same a-th order belief. Hence, both types induce the same $(a + 1)$-th order belief.
Now we can employ our recursive argument, starting at $p = k^* - 2$. For $p \in \{1, ..., k^* - 2\}$, assume that types $t_2^m[c_1', \bar{c}_1, (p + 1)a - 1]$ and $t_2^n[c_1', \bar{c}_1, (p + 1)a - 1]$ induce the same $((k^* - p - 1)a + 1)$-th order belief. Then types $t_1^m[c_1', \bar{c}_1, pa]$ and $t_1^n[c_1', \bar{c}_1, pa]$ induce the same $(k^* - p)a$-th order belief. Namely, from the beginning of the proof of this claim we have that types $t_i^m[c_1', \bar{c}_1, pa + k]$ and $t_i^n[c_1', \bar{c}_1, pa + k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., a - 2\}$ induce a probability one belief and moreover induce the same first-order belief. Therefore, they induce the same $(a-1)$-th order belief. Additionally, types $t_1^m[c_1', \bar{c}_1, (p+1)a-2]$ and $t_1^n[c_1', \bar{c}_1, (p+1)a-2]$ assign probability one to types that by assumption induce the same $((k^* - p)a + 1)$-th order belief. It follows then that types $t_1^m[c_1', \bar{c}_1, pa]$ and $t_1^n[c_1', \bar{c}_1, pa]$ induce the same $(k^* - p)a$-th order belief.

Now recall that for each $c_1', \bar{c}_1 \in C_1^\infty$, we have that

$$b_2[t_2^n[c_1', \bar{c}_1, pa - 1]](d^n[c_1', c_1'', pa], t_1^n[c_1', c_1'', pa]) = b_1^{\bar{c}_1}[c_1''] =$$
$$b_2[t_2^m[c_1', \bar{c}_1, pa - 1]](d^m[c_1', c_1'', pa], t_1^m[c_1', c_1'', pa])$$

Both types $t_2^m[c_1', \bar{c}_1, pa-1]$ and $t_2^n[c_1', \bar{c}_1, pa-1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same $(k^*-p)a$-th order belief. Hence, both types induce the same $((k^* - p)a + 1)$-th order belief.

Following the same argument, we can establish that types $t_1^m[c_1', c_1', 0]$ and $t_1^n[c_1', c_1', 0]$ induce the same k^*a-th order belief. From the above we know that types $t_2^m[c_1', c_1', a - 1]$ and $t_2^n[c_1', c_1', a - 1]$ induce the same $((k^* - 1)a + 1)$-th order belief. From the beginning of the proof of this claim we have that types $t_i^m[c_1', c_1', k]$ and $t_i^n[c_1', c_1', k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., a - 2\}$ induce a probability one belief and moreover induce the same first-order belief. Therefore, they induce

the same $(a-1)$-th order belief. Additionally, types $t_1^m[c_1', c_1', a-2]$ and $t_1^n[c_1', c_1', a-2]$ assign probability one to types that by the above recursive argument induce the same $((k^*-1)a+1)$ order belief. It follows then that types $t_1^m[c_1', c_1', 0]$ and $t_1^n[c_1', c_1', 0]$ induce the same k^*a-th order belief. This goes for each $c_1' \in C_1^\infty$.

Denote type $t_2^{n+1}[c_1, \bar{c}_1, k^*a-1]$ that results from our recursive backwards procedure but *before* constructing $\mathcal{M}^*$ by $\bar{t}_2^{n+1}[c_1, \bar{c}_1, k^*a-1]$. In contrast, let the same type that does result from constructing $\mathcal{M}^*$ still be denoted as $t_2^{n+1}[c_1, \bar{c}_1, k^*a-1]$. Now, we have for each $c_1, \bar{c}_1 \in C_1^\infty$

$$b_2[\bar{t}_2^{n+1}[c_1, \bar{c}_1, k^*a-1]](c_1', t_2^n[c_1', c_1', 0]) = b_2[t_2^{n+1}[c_1, \bar{c}_1, k^*a-1]](c_1', t_2^n[c_1', c_1', 0]), \forall c_1' \in C_1^\infty.$$

It thus follows that each such type $t_2^{n+1}[c_1, \bar{c}_1, k^*a-1]$ induces the same (k^*a+1)-th order belief in $\mathcal{M}^*$ as it did before $\mathcal{M}^*$ was constructed. All the remaining types in $\bigcup_{l \in \{n+1,...,m\}} T(l)$ remained unchanged when $\mathcal{M}^*$ was constructed: they induce exactly the same belief over choice-type combinations as before. As a result, all types in $\bigcup_{l \in \{n+1,...,m\}} T(l)$ induce at least the same (k^*a+1)-th order belief in $\mathcal{M}^*$ as before $\mathcal{M}^*$ was constructed.

In our backward construction procedure of types and choices, before creating $\mathcal{M}^*$, we constructed each $d^l[c_1, c_1', k]$ for each $l \in \{n+1, ..., m\}$, $k \in \{1, ..., k^*a-1\}$ and $c_1, c_1' \in C_1^\infty$ such that it is optimal given type $t_i^l[c_1, c_1', k]$. Now, we have that the maximum directly utility-relevant order of belief for any player is k^* and that each type $t_i^l[c_1, c_1', k]$ at least induces exactly the same (k^*a+1)-th order belief in $\mathcal{M}^*$ as it did before constructing $\mathcal{M}^*$. Hence, we also have in $\mathcal{M}^*$ that $d^l[c_1, c_1', k]$ is optimal given $t_i^l[c_1, c_1', k]$. This completes the proof of this claim.

Since each type in $\mathcal{M}^*$ only assigns positive probability to choice-type combinations $(d^l[c_1, c_1', k], t_i^l[c_1, c_1', k]])$ for $k \in \{0, 1, ..., k^*a-1\}$, each type only assigns positive probability to choice-type combinations where the choice is optimal given the type. Hence each type in $\mathcal{M}^*$ expresses 1-fold belief in rationality. Therefore also each type in $\mathcal{M}^*$ expresses common belief in rationality.

By our backward, recursive construction, we also have that type $t_1^m[c_1, c_1, 0]$ induces an a-th order expectation $b_1^{c_1}$. By construction of Step 1, choice c_1 is optimal given such a higher-order expectation. Hence we have constructed an epistemic model with a type that expresses common belief in rationality and is such that c_1 is optimal given that type.

In Step 1 we have shown that for every choice $c_1 \in C_1^\infty$ we can construct a partial epistemic model with a type that expresses on-path belief in rationality and that is such that choice c_1 is optimal. In Step 2 we showed that we are then also able to construct a finite, epistemic model with a type that expresses common belief in rationality and that is such that choice c_1 is optimal. This concludes the proof for Scenario (i).

Scenario (ii)

Step 1

Next consider scenario (ii) with $N_1 = \{a\}$ and $N_2 = \{z\}$, a, z odd. In this Step 1, we will construct a partial epistemic model.

For each choice $c_1 \in C_1^\infty$, fix an a-th order expectation $b_1^{c_1} \in \Delta(C_2^\infty)$ for which c_1 is optimal.[5]

[5]With a-th order expectation in this context we specifically refer to $marg_{C_2} e_1^a \in \Delta(C_2)$ where $e_1^a \in \Delta(W_1^{a-1} \times C_2)$.

The reason we can do so is as follows. From Lemma 1 we know that for each choice that is not strictly dominated in a decision problem, we can find a belief in that decision problem such that the relevant choice is optimal. The final reduced decision problem resulting from the IESDC-procedure leaves the choices in C_1^∞ for player 1 in the decision problem and the choices in C_2^∞ for player 2. That is, we have as a reduced decision problem after following through with the IESDC-procedure: $(C_1^\infty, C_2^\infty, v_i)$. By Lemma 1, we should then have that each choice in C_1^∞ is optimal for some a-th order expectation in $\Delta(C_2^\infty)$. Then, for each $c_1 \in C_1^\infty$, construct a type $t_1^{c_1}$. Similarly, also for each choice $c_2 \in C_2^\infty$, fix a z-th order expectation $b_2^{c_2} \in \Delta(C_1^\infty)$ for which c_2 is optimal. Again, we can do so for the reasons explained above, but then from player 2's perspective. Then, for each $c_2 \in C_2^\infty$, construct a type $t_2^{c_2}$.

Take for each $t_1^{c_1}$ a sequence of types $(t_1^{c_1}, t_2^{c_1,1}, ..., t_1^{c_1,a-1})$. Then, let us have in each such sequence that type $t_1^{c_1}$ assigns probability one to type $t_2^{c_1,1}$ if $a > 1$, and type $t_i^{c_1,n}$ probability one to type $t_j^{c_1,n+1}$ for each $n \in \{1, 2, ..., a-2\}$ and with $i \in \{1,2\}$ and $j \neq i$. Note that if $a = 1$, we treat type $t_1^{c_1}$ such that $t_1^{c_1} = t_1^{c_1,a-1}$. Similarly, take for each $t_2^{c_2}$ a sequence of types $(t_2^{c_2}, t_1^{c_2,1}, ..., t_2^{c_2,z-1})$. Then, for each c_2, let us have in each such sequence that type $t_2^{c_2}$ assigns probability one to type $t_1^{c_2,1}$ if $z > 1$, and type $t_i^{c_2,n}$ probability one to type $t_j^{c_2,n+1}$ for each $n \in \{1, 2, ..., z-2\}$. Again, if $z = 1$, we treat type $t_2^{c_2}$ such that $t_2^{c_2} = t_2^{c_2,z-1}$.

Finally, for each $c_1 \in C_1^\infty$ define type $t_1^{c_1,a-1}$ to be such that, for each $c_2' \in C_2^\infty$,

$$b_1[t_1^{c_1,a-1}](c_2', t_2^{c_2''}) := \begin{cases} b_1^{c_1}(c_2'), \text{ if } c_2' = c_2'' \\ 0, \text{ otherwise.} \end{cases}$$

So we have that the distribution over choices induced by type $t_1^{c_1,a-1}$ is equal to the distribution over choices represented by the expectation $b_1^{c_1}$. From type $t_1^{c_1}$ there follows a sequence of probability one beliefs up to type $t_1^{c_1,a-1}$. The a-th order expectation induced by type $t_1^{c_1}$ is thus equal to $b_1^{c_1}$. Hence, choice c_1 is optimal given type $t_1^{c_1}$.

Similarly, for each $c_2 \in C_2^\infty$ define type $t_2^{c_2,z-1}$ to be such that, for each $c_1' \in C_1^\infty$

$$b_2[t_2^{c_2,z-1}](c_1', t_1^{c_1''}) := \begin{cases} b_2^{c_2}(c_1'), \text{ if } c_1' = c_1'' \\ 0, \text{ otherwise.} \end{cases}$$

So we have that the distribution over choices induced by type $t_2^{c_2,z-1}$ is equal to the distribution over choices represented by the expectation $b_2^{c_2}$. From type $t_2^{c_2}$ there follows a sequence of probability one beliefs up to type $t_2^{c_2,z-1}$. Similarly as before, then type $t_2^{c_2}$ then induces a z-th order expectation that is equal to $b_2^{c_2}$. Then choice c_2 is optimal given type $t_2^{c_2}$.

Call the resulting partial epistemic model $\mathcal{M}$. By construction, we have for each $c_1 \in C_1^\infty$ that c_1 is optimal given $t_1^{c_1}$, and we have for each $c_2 \in C_2^\infty$ that c_2 is optimal for $t_2^{c_2}$. Moreover, each $t_1^{c_1}$ also expresses on-path belief in rationality. This is because the type $t_1^{c_1,a-1}$ only assigns positive probability to choice type pairs $(c_2', t_2^{c_2'})$. In these pairs the choice is optimal for the type by construction. For similar reasons, each type $t_2^{c_2}$ also expresses on-path belief in rationality.

Step 2

In Step 1 we have shown that for each choice $c_1 \in C_1^\infty$ we can always construct a *partial* epistemic model with a type $t_1^{c_1}$ for which c_1 is optimal and that expresses on-path belief in rationality. In Step 2 we will now do the following. We will show that if there exists a belief hierarchy expressing

on-path belief in rationality for which c_1 is optimal, then there is also a belief hierarchy expressing common belief in rationality for which c_1 is optimal.

Consider a partial epistemic model $\mathcal{M}$ as constructed in Step 1 with a type $t_1^{c_1}$ that expresses on-path belief in rationality and for which c_1 is optimal. By means of a backward, recursive procedure we transform this partial epistemic model such that we get to a new, complete epistemic model that now includes a completed type $t_1^m[c_1, 0]$ that expresses common belief in rationality and induces the same a-th order expectation as type $t_1^{c_1}$ does. The recursive procedure in each iteration defines combinations of choices *and* types at the same time.

The recursive procedure is as follows.

Iteration 0: For each choice $c_1 \in C_1^\infty$, define

$$d^0[c_1, 0] := c_1.$$

Moreover, for each choice $c_1 \in C_1^\infty$ and each $k \in \{1, 2, ..., a-1\}$, define $d^0[c_1, k]$ randomly. So

$$d^0[c_1, k] := c', \text{ for } \textit{some } c' \in C_1^\infty \text{ if } k \text{ is even or } \textit{some } c' \in C_2^\infty \text{ if } k \text{ is odd.}$$

Now, take a sequence of types $(t_1^0[c_1, 0], t_2^0[c_1, 1], ..., t_1^0[c_1, a-1])$ for every choice $c_1 \in C_1^\infty$. Define type $t_1^0[c_1, 0]$ to be such that

$$b_1[t_1^0[c_1, 0]] := (d^0[c_1, 1], t_2^0[c_1, 1]).$$

Then, define for each $k \in \{1, 2, ..., a-2\}$ type $t_i^0[c_1, k]$ with $i \in \{1, 2\}$ to be such that

$$b_i[t_i^0[c_1, k]] := (d^0[c_1, k+1], t_j^0[c_1, k+1]).$$

Finally, we define for each choice $c_1 \in C_1^\infty$ type $t_1^0[c_1, a-1]$ to be such that

$$b_1[t_1^0[c_1, a-1]](c_2', t_2^0[c_2', 0]) := b_1^{c_1}(c_2'), \forall c_2' \in C_2^\infty.$$

This implies that type $t_1^0[c_1, 0]$ induces exactly the same a-th order expectation as type $t_1^{c_1}$ did in Step 1. So for Iteration 0 we essentially take a copy of the epistemic model created in Step 1, but fill in the beliefs that were still incomplete from this step.

We do a similar thing for each choice c_2. For each choice $c_2 \in C_2^\infty$, define

$$d^0[c_2, 0] := c_2.$$

Moreover, for each choice $c_2 \in C_2^\infty$ and each $k \in \{1, 2, ..., z-1\}$, define $d^0[c_2, k]$ randomly. So

$$d^0[c_2, k] := c', \text{ for } \textit{some } c' \in C_2^\infty \text{ if } k \text{ is even or } \textit{some } c' \in C_1^\infty \text{ if } k \text{ is odd.}$$

Now, take a sequence of types $(t_2^0[c_2, 0], t_1^0[c_2, 1], ..., t_2^0[c_2, z-1])$ for every choice $c_2 \in C_2^\infty$. Define type $t_2^0[c_2, 0]$ to be such that

$$b_2[t_2^0[c_2, 0]] := (d^0[c_2, 1], t_1^0[c_2, 1]).$$

Then, define for each $k \in \{1, 2, ..., z-2\}$ type $t_i^0[c_2, k]$ with $i \in \{1, 2\}$ to be such that

$$b_i[t_i^0[c_2, k]] := (d^0[c_2, k+1], t_j^0[c_2, k+1]).$$

Finally, we define for each choice $c_2 \in C_2^\infty$ type $t_2^0[c_2, z-1]$ to be such that

$$b_2[t_2^0[c_2, z-1]](c_1', t_1^0[c_1', 0]) := b_2^{c_2}(c_1'), \forall c_1' \in C_1^\infty.$$

Iteration n$\geqslant$ 1 : We define for each choice $c_2 \in C_2^\infty$ type $t_2^n[c_2, z-1]$ to be such that

$$b_2[t_2^n[c_2, z-1]](c_1', t_1^{n-1}[c_1', 0]) := b_2^{c_2}(c_1'), \forall c_1' \in C_1^\infty.$$

We also define

$$d^n[c_2, z-1] := c_2', \text{ with } c_2' \text{ optimal given the type } t_2^n[c_2, z-1].$$

Define recursively for each *odd* $k \in \{1, 2, ..., z-2\}$ starting at $k = z-2$, type $t_1^n[c_2, k]$ that is such that

$$b_1[t_1^n[c_2, k]] := (d^n[c_2, k+1], t_2^n[c_2, k+1]).$$

Second, also define

$$d^n[c_2, k] := c_1', \text{ with } c_1' \text{ optimal given the a-th order expectation induced by } t_1^n[c_2, k].$$

Third, define type $t_2^n[c_2, k-1]$ to be such that

$$b_2[t_2^n[c_2, k-1]] := (d^n[c_2, k], t_1^n[c_2, k]).$$

Fourth, also define

$$d^n[c_2, k-1] := c_2'', \text{ with } c_2'' \text{ optimal given the type } t_2^n[c_2, k-1].$$

Finally, for each choice $c_2 \in C_2^\infty$ define type $t_2^n[c_2, 0]$ to be such that

$$b_2[t_2^n[c_2, 0]] := (d^n[c_2, 1], t_1^n[c_2, 1]),$$

and define

$$d^n[c_2, 0] := c_2.$$

Next, we do exactly the same thing for choices c_1. We define for each choice $c_1 \in C_1^\infty$ type $t_1^n[c_1, a-1]$ to be such that

$$b_1[t_1^n[c_1, a-1]](c_2', t_2^n[c_2', 0]) := b_1^{c_1}(c_2'), \forall c_2' \in C_2^\infty.$$

For each choice $c_1 \in C_1^\infty$, we then also define

$$d^n[c_1, a-1] := c_1', \text{ with } c_1' \text{ optimal given the type } t_1^n[c_1, a-1].$$

Now, for each choice $c_1 \in C_1^\infty$, define recursively for each *odd* $k \in \{1, 2, ..., a-2\}$ starting at $k = a-2$, type $t_2^n[c_1, k]$ that is such that

$$b_2[t_2^n[c_1, k]] := (d^n[c_1, k+1], t_1^n[c_1, k+1]).$$

Second, also define

$$d^n[c_1, k] := c_2', \text{ with } c_2' \text{ optimal given the a-th order expectation induced by } t_2^n[c_1, k].$$

Third, define type $t_1^n[c_1, k-1]$ to be such that

$$b_1[t_1^n[c_1, k-1]] := (d^n[c_1, k], t_2^n[c_1, k]).$$

Fourth, also define

$$d^n[c_1, k-1] := c_1'', \text{ with } c_1'' \text{ optimal given the type } t_1^n[c_1, k-1].$$

Finally, for each choice $c_1 \in C_1^\infty$ define type $t_1^n[c_1, 0]$ to be such that

$$b_1[t_1^n[c_1, 0]] := (d^n[c_1, 1], t_2^n[c_1, 1]),$$

and define

$$d^n[c_1, 0] := c_1.$$

We have that C_1^∞ and C_2^∞ are finite sets. Moreover, a and z are both finite orders of belief. Hence, there are iterations m and n with $m > n$ such that

$$d^m[c_1, k] = d^n[c_1, k] \text{ and } d^m[c_2, l] = d^n[c_2, l],$$

for each $c_1 \in C_1^\infty$ and $k \in \{0, 1, ..., a-1\}$, and $c_2 \in C_1^\infty$ and $l \in \{0, 1, ..., z-1\}$ respectively. When we find such iterations m and n, we stop the procedure.

Next, we create the epistemic model $\mathcal{M}^*$ from the types we have constructed in our recursive procedure. Define $T_1(l) := \{t_1^l[c_1, k] : c_1 \in C_1^\infty, k \in \{2, ..., a-1\} \text{ even}\} \cup \{t_1^l[c_2, k] : c_2 \in C_2^\infty, k \in \{1, ..., z-2\} \text{ odd}\}$ and $T_2(l) := \{t_2^l[c_2, k] : c_2 \in C_2^\infty, k \in \{2, ..., z-1\} \text{ even}\} \cup \{t_2^l[c_1, k] : c_1 \in C_1^\infty, k \in \{1, ..., a-2\} \text{ odd}\}$. Then, let $T(l) := T_1(l) \cup T_2(l)$. Do this for every $l \in \{n, ..., m\}$.

In $T(n+1)$ specifically, we re-define for each $c_2 \in C_2^\infty$ the type $t_2^{n+1}[c_2, z-1]$. Re-define each such type $t_2^{n+1}[c_2, z-1]$ to be such that

$$b_2[t_2^{n+1}[c_2, z-1]](c_1', t_1^m[c_1', 0]) := b_2^{c_2}(c_1'), \forall c_1' \in C_1^\infty.$$

So instead of assigning positive probability to types in $T(n)$, each type $t_2^{n+1}[c_2, z-1]$ now assigns positive probability to types in $T(m)$. Then define $\mathcal{M}^* := (\bigcup_{l \in \{n+1, ..., m\}} T_i(l), b[t_i])_{i \in \{1,2\}}$.

We will show that each type in $\mathcal{M}^*$ expresses common belief in rationality. We will do so in steps.

First we note that choice c_1 is optimal for type $t_1^l[c_1, 0]$, for each $c_1 \in C_1^\infty$ and each $l \in \{n+1, ..., m\}$ in $\mathcal{M}^*$. Namely, from type $t_1^l[c_1, 0]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_1^l[c_1, 0], t_2^l[c_1, 1], ..., t_2^l[c_1, a-2])$. This sequence of probability one beliefs ends at type $t_1^l[c_1, a-1]$. Note that if $a = 1$ we treat $t_1^l[c_1, 0]$ as if $t_1^l[c_1, 0] = t_1^l[c_1, a-1]$. By definition, we have that

$$marg_{C_2^\infty} b_1[t_1^l[c_1, a-1]] = b_1^{c_1}.$$

The a-th order expectation induced by type $t_1^l[c_1, 0]$ is thus equal to $b_1^{c_1}$. We constructed $b_1^{c_1}$ such that c_1 is optimal given $b_1^{c_1}$. Hence c_1 is optimal given type $t_1^l[c_1, 0]$. This goes for every

$l \in \{n+1, ..., m\}$. For similar reasons, we have that for each $l \in \{n+1, ..., m\}$ and each $c_2 \in C_2^\infty$ that $t_2^l[c_2, 0]$ induces the same z-th order expectation as $b_2^{c_2}$ does. Hence, c_2 is optimal given type $t_2^l[c_2, 0]$.

Second, we can also show the following is true.

Claim 3. *Consider the epistemic model $\mathcal{M}^*$. For each $l \in \{n+1, ..., m\}$, each $c_1 \in C_1^\infty$ and each $k \in \{1, 2, ..., a-1\}$, each choice $d^l[c_1, k]$ is optimal given the type $t_i^l[c_1, k]$. Moreover, for each $l \in \{n+1, ..., m\}$, each $c_2 \in C_2^\infty$ and each $k \in \{1, 2, ..., z-1\}$, each choice $d^l[c_2, k]$ is optimal given the type $t_i^l[c_2, k]$.*

Proof of claim. For each $k \in \{0, 2, ..., a-2\}$ and each $c_1' \in C_1^\infty$ we have by construction that

$$b_i[t_i^n[c_1', k]](d^n[c_1', k+1], t_j^n[c_1', k+1]) = 1 = b_i[t_i^m[c_1', k]](d^m[c_1', k+1], t_j^m[c_1', k+1])$$

with $d^m[c_1', k+1] = d^n[c_1', k+1]$. We similarly have for each $k \in \{0, 2, ..., z-2\}$ and each $c_2' \in C_2^\infty$ that

$$b_i[t_i^n[c_2', k]](d^n[c_2', k+1], t_j^n[c_2', k+1]) = 1 = b_i[t_i^m[c_2', k]](d^m[c_2', k+1], t_j^m[c_2', k+1])$$

with $d^m[c_2', k+1] = d^n[c_2', k+1]$. Additionally, we have that

$$b_1[t_1^n[c_1', a-1]](c_2'', t_2^{n-1}[c_2'', 0]) = b_1^{c_1'}(c_1'') = b_1[t_1^m[c_1', a-1](c_2'', t_2^{m-1}[c_2'', 0]),$$

for each $c_2'' \in C_1^\infty$. And we have that

$$b_2[t_2^n[c_2', z-1]](c_1'', t_1^{n-1}[c_1'', 0]) = b_2^{c_2'}(c_1'') = b_2[t_2^m[c_2', z-1](c_1'', t_1^{m-1}[c_1'', 0]),$$

for each $c_1'' \in C_1^\infty$.

Then for each $c_1' \in C_1^\infty$ the pair of types $t_1^n[c_1', 0]$ and $t_1^m[c_1', 0]$ induce the same $(z+a)$-th order *belief*. To see this, we can first note that for each $c_2'' \in C_2^\infty$ the pair of types $t_2^n[c_2'', 0]$ and $t_2^m[c_2'', 0]$ induce the same z-th order belief. Namely, from the beginning of the proof of this claim we have that types $t_i^m[c_2'', k]$ and $t_i^n[c_2'', k]$ for each $k \in \{0, 1, ..., z-2\}$ induce a probability one belief. Moreover, the first-order belief induced by type $t_i^m[c_2'', k]$ for each $k \in \{0, 1, ..., z-1\}$ is equal to the first-order belief induced by $t_i^n[c_2'', k]$. As a result, types $t_2^m[c_2'', 0]$ and $t_2^n[c_2'', 0]$ in fact induce the same z-th order belief.

Recall that for each $c_1' \in C_1^\infty$ we have that

$$b_1[t_1^n[c_1', a-1]](c_2'', t_2^{n-1}[c_2'', 0]) = b_1^{c_1'}(c_1'') = b_1[t_1^m[c_1', a-1](c_2'', t_2^{m-1}[c_2'', 0]),$$

for each $c_2'' \in C_1^\infty$. Both types $t_1^m[c_1', a-1]$ and $t_1^n[c_1', a-1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same z-th order belief as established before. It follows that for each $c_1' \in C_1^\infty$ types $t_1^n[c_1', a-1]$ and $t_1^m[c_1', a-1]$ induce the same $(z+1)$-th order belief.

From the beginning of the proof of this claim we have that types $t_i^m[c_1', k]$ and $t_i^n[c_1', k]$ for each $k \in \{0, 1, ..., a-2\}$ induce a probability one belief.. These probability one beliefs end at types $t_1^m[c_1', a-1]$ and $t_1^n[c_1', a-1]$ respectively. We know these types induce the same $(z+1)$-th order expectation. Moreover, the first-order belief induced by type $t_i^m[c_1', k]$ for each $k \in \{0, 1, ..., a-2\}$

is equal to the first-order belief induced by $t_i^n[c_1', k]$. Taken together, types $t_1^m[c_1', 0]$ and $t_1^n[c_1', 0]$ in fact induce the same $(z + a)$-th order belief.

Denote type $t_2^{n+1}[c_2, z-1]$ that results from our recursive, backward procedure but *before* constructing $\mathcal{M}^*$ by $\bar{t}_2^{n+1}[c_2, z-1]$. Now, we have that for each $c_2 \in C_2^\infty$

$$b_2[\bar{t}_2^{n+1}[c_2, z-1]](c_1', t_1^n[c_1', 0]) = b_2[t_2^{n+1}[c_2, z-1]](c_1', t_1^m[c_1', 0]), \forall c_1' \in C_1^\infty.$$

It follows that each such type $t_2^{n+1}[c_2, z-1]$ induces the same $(z+a+1)$-th order belief in $\mathcal{M}^*$ as it did before $\mathcal{M}^*$ was constructed. All the remaining types in $\bigcup_{l \in \{n+1,...,m\}} T(l)$ remained unchanged when $\mathcal{M}^*$ was constructed: they induce exactly the same belief over choice-type combinations as before. As a result, all types in $\bigcup_{l \in \{n+1,...,m\}} T(l)$ induce at least the same $(z+a+1)$-th order belief in $\mathcal{M}^*$ as before $\mathcal{M}^*$ was constructed.

In our backward construction procedure of types and choices, before creating $\mathcal{M}^*$, we constructed $d^l[c_1, k]$ for each $l \in \{n+1, ..., m\}$, $k \in \{1, 2, ..., a-1\}$ and $c_1 \in C_1^\infty$ such that it is optimal given type $t_i^l[c_1, k]$. Similarly, we constructed $d^l[c_1, k]$ for each $l \in \{n+1, ..., m\}$, $k \in \{1, 2, ..., z-1\}$ and $c_2 \in C_2^\infty$ such that it is optimal given type $t_i^l[c_2, k]$. Now, we have that the maximum order of belief in which either of the players' utility is variable is $max(N_1 \cup N_2)$, which is either a or z. We also have that types $t_i^l[c_1, k]$ and $t_i^l[c_2, k]$ induce exactly the same $(z+a+1)$-th order belief in $\mathcal{M}^*$ as before $\mathcal{M}^*$ was constructed. Hence, we also have in $\mathcal{M}^*$ that $d^l[c_1, k]$ is optimal given $t_i^l[c_1, k]$ and that $d^l[c_2, k]$ is optimal given $t_i^l[c_2, k]$. This completes the proof of this claim.

Since each type in the epistemic model $\mathcal{M}^*$ only assigns positive probability to choice-type combinations $(d^l[c_1, k], t_i^l[c_1, k])$ for $k \in \{0, 1, ..., a-1\}$ or $(d^l[c_2, k], t_i^l[c_2, k])$ for $k \in \{0, 1, ..., z-1\}$, each type only assigns positive probability to choice-type combinations where the choice is optimal given the type. Hence each type in $\mathcal{M}^*$ expresses 1-fold belief in rationality. Therefore also each type in $\mathcal{M}^*$ expresses common belief in rationality.

By our backward, recursive construction, we also have that type $t_1^m[c_1, 0]$ induces an a-th order expectation $b_1^{c_1}$. By construction of Step 1, choice c_1 is optimal given such a higher-order expectation. Hence we have constructed an epistemic model with a type that expresses common belief in rationality and is such that c_1 is optimal given that type.

In Step 1 we have shown that for every choice $c_1 \in C_1^\infty$ we can construct a partial epistemic model with a type that expresses on-path belief in rationality and that is such that choice c_1 is optimal. In Step 2 we showed that we are then also able to construct a finite, epistemic model with a type that expresses common belief in rationality and that is such that choice c_1 is optimal. This concludes the proof for Scenario (ii).

Scenario (iii)
Step 1:
Finally, consider scenario (iii). This corresponds to case (iii) of Theorem 2. Here we have that $N_1 = \{a, b, ..., x\}$ consists of (possibly multiple) odd orders and that N_2 is of a single, even order such that in the resulting causality diagram for player 1 there are no overlapping paths. In this Step 1, we will construct a partial epistemic model. By *partial* we mean we only completely specify the beliefs for some particular types.

For each choice $c_1 \in C_1^\infty$, fix an expectation $b_1^{c_1}$ which is a probability distribution over the product-

space $C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty}$ and for which c_1 is optimal. The reason we can do so is as follows. From Lemma 1 we know that for each choice that is not strictly dominated in a decision problem, we can find a belief in that decision problem such that the relevant choice is optimal. The final reduced problem decision resulting from the IESDC-procedure leaves the choices in C_2^∞ for player 2 in the decision problem. By Lemma 1, we should then have that each choice in C_1^∞ is optimal for some belief in $\Delta(C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty})$. Each letter in the superscripts of the product space $C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty}$ refers to an order of belief in N_1.

Then, fix a type $t_1^{c_1}[c_1]$ for choice c_1.

The lowest order in N_1 is order a. Let $N_2 = \{z\}$ and take $a + z$. For each remaining order $p \in N_1$, subtract a multiple $n \in \mathbb{N}$ of z from order p such that $p - n \cdot z \in \{a, ..., a + z\}$. Call this order a^p. Note here that, by how case (iii) in Theorem 2 is defined, for any orders $b, c \in N_1$ with $b, c \neq a$, we have that $b - n \cdot z \neq c - m \cdot z$, for any combination of n, m. Hence, $a^b \neq a^c$ for any two different orders $b, c \in N_1$.

Take some combination of choices for player 2 $(c^a, c^b, ..., c^x) \in supp(b_1^{c_1})$. For each order $p \in N_1$ we do the following: take choice c^p in $C_2^{p,\infty}$. Let

$$b_2^{c^{p-z}} := c^p$$

be the z-th order expectation for player 2 that puts probability one on c^p.[6] Then there is a choice c^{p-z} in $C_2^{p-z,\infty}$ such that c^{p-z} is optimal given $b_2^{c^{p-z}}$. This follows from the construction of the IESDC-procedure. Namely, every choice $c_2 \in C_2$ that is optimal for some belief $b_2 \in \Delta(C_2^\infty)$ is in C_2^∞.

Next, for each $n \geqslant 1$, up until we have $p - n \cdot z \in \{a, ..., a + z\}$, we can do the same as we did for choice c^p. Take choice $c^{p-(n-1)z}$. Let

$$b_2^{c^{p-nz}} := c^{p-(n-1)z},$$

be the z-th order expectation for player 2 that puts probability one on $c^{p-(n-1)z}$. Then following the same argument as before there is a choice c^{p-nz} in $C_2^{p-nz,\infty}$ such that c^{p-nz} is optimal given $b_2^{c^{p-nz}}$. We can do this for any $p > a + z$ with $p \in N_1$, up until we have the choice c^{a^p} in $C_2^{a^p,\infty}$.

Each choice c^{p-nz}, given any p and any n, is a choice in C_2^∞. For these choices, we fix the z-th order expectation $b_2^{c^{p-nz}}$ we just constructed before. For all the remaining choices c_2 in C_2^∞, we fix some z-th order expectation b^{c_2} such that c_2 is optimal given $b_2^{c_2}$. Again, we can do so by Lemma 1.

The next step is to move on to the construction of types. For *each* combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$ that results from the construction above, create a type $t_2^{\bar{c}}[\bar{c}]$. We specifically say '*each*', as the support of $b_1^{c_1}$ may include multiple combinations of choices $(c^a, c^b, ..., c^x)$.

For each combination of choices $\bar{c}$, take a sequence of types $(t_2^{\bar{c}}[\bar{c}], t_1^{\bar{c},1}[\bar{c}], ..., t_1^{\bar{c},z-1}[\bar{c}])$. Then in each such sequence let us have that type $t_2^{\bar{c}}[\bar{c}]$ assigns probability one to type $t_1^{\bar{c},1}[\bar{c}]$. Also in each such sequence, let us have, for each $n \in \{1, 2, ..., z-1\}$ that type $t_i^{\bar{c},n}[\bar{c}]$ with $i \in \{i, j\}$ assigns probability one to type $t_j^{\bar{c},n+1}[\bar{c}]$ with $j \neq i$. Additionally, for each $k = a^p - a - 1$, define type $t_1^{\bar{c},k}[\bar{c}]$ such that

$$b_1[t_1^{\bar{c},k}[\bar{c}]] := (c^{a^p}, t_2^{\bar{c},k}[\bar{c}]).$$

Finally, we need to specify the belief that type $t_1^{\bar{c},z-1}[\bar{c}]$ in the sequence induces. To this end, first construct for each $\bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ and each $\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ a type $t_2^{\bar{c},z}[\bar{c}']$. Then,

[6] With z-th order expectation in this context we specifically refer to $marg_{C_2} e_2^z \in \Delta(C_2)$ where $e_2^z \in \Delta(W_2^{z-1} \times C_2)$.

consider the joint probability distribution $b^{\bar{c}} \in \Delta(C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x})$. That is,

$$b^{\bar{c}}(\bar{c}') := b_2^{c^a}(c^{a\prime}) \cdot b_2^{c^{a^b}}(c^{a^b\prime}) \cdot ... \cdot b_2^{c^{a^x}}(c^{a^x\prime}),\ \forall \bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}.$$

Then, define type $t_1^{\bar{c},z-1}[\bar{c}]$ to be such that, for each $\bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$

$$b_1[t_1^{\bar{c},z-1}[\bar{c}]](c^{a\prime}, t_2^{\bar{c},z}[\bar{c}'']) := \begin{cases} b^{\bar{c}}(\bar{c}'), \text{ if } \bar{c}' = \bar{c}'' \\ 0, \text{ otherwise.} \end{cases}$$

We create such sequences of types for each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$.

Now, we follow the same construction another $k^* - 2$ times, where $k^* = max(N_1 \cup N_2)$. For each $y \in \{1, ..., k^* - 2\}$, do the following: For each $\bar{c}, \bar{c'} \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$ take a sequence of types $(t_2^{\bar{c},yz}[\bar{c}'], ..., t_1^{\bar{c},(y+1)z-1}[\bar{c}'])$. Then, let us have in each such sequence that type $t_2^{\bar{c},yz}[\bar{c}']$ assigns probability one to type $t_1^{\bar{c},yz+1}[\bar{c}']$, and that type $t_i^{\bar{c},yz+n}[\bar{c}']$ with $i \in \{1, 2\}$ assigns probability one to type $t_j^{\bar{c},yz+n+1}[\bar{c}']$ with $j \neq i$, for each $n \in \{1, ..., z-2\}$. Additionally, for each $k = a^p - a - 1$, define type $t_1^{\bar{c},yz+k}[\bar{c}']$ to be such that

$$b_1[t_1^{\bar{c},yz+k}[\bar{c}']] := (c^{a^p\prime}, t_2^{\bar{c},yz+k+1}[\bar{c}']),$$

with $\bar{c}' = (c^{a\prime}, ..., c^{a^x\prime})$. Then, for each $\bar{c}, \bar{c}^* \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$ construct a type $t_2^{\bar{c},(y+1)z}[\bar{c}^*]$. Then define type $t_1^{\bar{c},(y+1)z-1}[\bar{c}']$ to be such that for each $\bar{c}^* = (c^{a^*}, c^{a^b*}, ..., c^{a^x*}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$

$$b_1[t_1^{\bar{c},(y+1)z-1}[\bar{c}']](c^{a*}, t_2^{\bar{c},(y+1)z}[\bar{c}'']) := \begin{cases} b^{\bar{c}'}(\bar{c}^*), \text{ if } \bar{c}^* = \bar{c}'' \\ 0, \text{ otherwise.} \end{cases}$$

We do this for every $y \in \{1, ..., k^* - 2\}$.

Finally, we consider the case $y = k^* - 1$. For each $\bar{c}, \bar{c'} \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$ take a sequence of types $(t_2^{\bar{c},(k^*-1)z}[\bar{c}'], ..., t_1^{\bar{c},k^*z-1}[\bar{c}'])$. Then, let us have in each such sequence that type $t_2^{\bar{c},(k^*-1)z}[\bar{c}']$ assigns probability one to type $t_1^{\bar{c},(k^*-1)z+1}[\bar{c}']$, and that type $t_i^{\bar{c},(k^*-1)z+n}[\bar{c}']$ with $i \in \{1, 2\}$ assigns probability one to type $t_j^{\bar{c},(k^*-1)z+n+1}[\bar{c}']$ with $j \neq i$, for each $n \in \{1, ..., z-2\}$. Additionally, for each $k = a^p - a - 1$, define type $t_1^{\bar{c},(k^*-1)z+k}[\bar{c}']$ to be such that

$$b_1[t_1^{\bar{c},(k^*-1)z+k}[\bar{c}']] := (c^{a^p\prime}, t_2^{\bar{c},(k^*-1)z+k+1}[\bar{c}']),$$

with $\bar{c}' = (c^{a\prime}, ..., c^{a^x\prime})$. Then define type $t_1^{\bar{c},k^*z-1}[\bar{c}']$ to be such that for each $\bar{c}^* = (c^{a^*}, c^{a^b*}, ..., c^{a^x*}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$

$$b_1[t_1^{\bar{c},k^*z-1}[\bar{c}']](c^{a*}, t_2^{\bar{c}''}[\bar{c}'']) := \begin{cases} b^{\bar{c}'}(\bar{c}^*), \text{ if } \bar{c}^* = \bar{c}'' \\ 0, \text{ otherwise.} \end{cases}$$

We create such k^* sequences of z types for each combination of choices $\bar{c} = (c^a, ..., c^{a^x}0 \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$. All these types together form a partial epistemic model. Call this partial epistemic model $\bar{\mathcal{M}}$.

Extend this partial epistemic model in the following way. Let type $t_1^{c_1}[c_1]$ we fixed at the beginning be at the start of the following sequence of types: $(t_1^{c_1}[c_1], t_2^{c_1,1}[c_1], ..., t_1^{c_1,a-1}[c_1])$. Let type $t_1^{c_1}[c_1]$ be such that it assigns probability one to type $t_2^{c_1,1}[c_1]$ and let type $t_i^{c_1,n}[c_1]$ be such that it assigns probability one to type $t_j^{c_1,n+1}[c_1]$, for each $n \in \{1, 2, ..., a-2\}$.

Now, we have that from each combination of choices $(c^a, c^b, ..., c^x)$ we derive a single combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$. Then, for each combination of choices $(c^a, c^b, ..., c^x)$ and the combination of choices $\bar{c}$ that is derived from it, define type $t_1^{c_1,a-1}[c_1]$ to be such that

$$b_1[t_1^{c_1,a-1}[c_1]](c^a, t_2^{\bar{c}'}[\bar{c}']) := \begin{cases} b_1^{c_1}(c^a, c^b, ..., c^x), \text{ if } \bar{c}' = \bar{c}, \\ 0, \text{ otherwise.} \end{cases}$$

As will be explained below, then type $t_1^{c_1}[c_1]$ expresses on-path belief in rationality in this partial epistemic model. We do so by first showing that each type $t_2^{\bar{c}}[\bar{c}]$ expresses z-fold belief in rationality, 2z-fold belief in rationality, and so on. Additionally, we show that this type expresses $(a^p - a)$-fold belief in rationality, $z + (a^p - a)$-fold belief in rationality and so on, for every order a^p.

First, it is clear that type $t_2^{\bar{c}}[\bar{c}]$ for any $\bar{c} \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x}$ expresses z-fold belief in rationality by construction. Namely, for any $\bar{c}' = (c^{a\prime}, ..., c^{a^x\prime})$, type $t_1^{\bar{c},2z-1}[\bar{c}']$ is such that its distribution over choices in C_2^{∞} in the belief it induces is exactly equal to $b^{\bar{c}'}$, which makes choice $c^{a\prime}$ optimal by construction. Since from type $t_1^{\bar{c},z}[\bar{c}']$ a sequence of $z-1$ probability one beliefs is induced that ends up at type $t_1^{\bar{c},2z-1}[\bar{c}']$, it follows that choice $c^{a\prime}$ is optimal given type $t_2^{\bar{c}}[\bar{c}']$. We have in the sequence $(t_2^{\bar{c}}[\bar{c}], ..., t_1^{\bar{c},z-1}[\bar{c}])$ that type $t_1^{\bar{c},z-1}[\bar{c}]$ is constructed such that it only assigns positive probability to choice-type combinations $(c^{a\prime}, t_2^{\bar{c},z}[\bar{c}']$ where the choice is optimal given such type. Hence type $t_2^{\bar{c}}[\bar{c}]$ expresses z-fold belief in rationality.

Similarly, type $t_2^{\bar{c}}[\bar{c}]$ expresses $(a^p - a)$-fold belief in rationality for each order a^p. To see this, first note that from type $t_2^{\bar{c},a^p-a}[\bar{c}]$ there follows a sequence of probability one beliefs up to type $t_1^{\bar{c},z-1}[\bar{c}]$. Second, type $t_1^{\bar{c},z-1}[\bar{c}]$ induces a belief whose distribution over types is such that it is equal to the distribution that $b^{\bar{c}}$ has over $C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$. In different terms, the belief $b_1[t_1^{\bar{c},z-1}[\bar{c}]]$ assigns a probability to type $t_2^{\bar{c},z}[\bar{c}']$ that is equal to the probability that $b^{\bar{c}} = b_2^{c^a} \times ... \times b_2^{c^{a^p}} \times ... \times b_2^{c^{a^x}}$ assigns to $\bar{c}' = (c^{a\prime}, ..., c^{a^p\prime}, ..., c^{a^x\prime})$. Third, by construction, from type $t_2^{\bar{c},z}[\bar{c}']$ there follows a sequence of $(a^p - a)$ probability one beliefs. The $(a^p - a)$-th type in this sequence assigns probability one to specifically the choice-type combination $(c^{a^p\prime}, t_2^{\bar{c},a^p-a}[\bar{c}'])$. Taken together then, type $t_2^{\bar{c},a^p-a}[\bar{c}]$ induces a z-th order expectation that is equal to $b_2^{c^{a^p}}$. By construction choice c^{a^p} is optimal given $b_2^{a^p}$. Hence choice c^{a^p} is optimal given type $t_2^{\bar{c},a^p-a}[\bar{c}]$. Since from type $t_2^{\bar{c}}[\bar{c}]$ there follows a sequence of probability one beliefs up to type $t_1^{\bar{c},a^p-a-1}[\bar{c}]$ and $b_1[t_1^{\bar{c},a^p-a-1}[\bar{c}]](c^{a^p}, t_2^{\bar{c},a^p-a}[\bar{c}]) = 1$, we have that type $t_2^{\bar{c}}[\bar{c}]$ expresses $(a^p - a)$-fold belief in rationality.

Following exactly the same argument, we have that type $t_2^{\bar{c},yz}[\bar{c}']$ for each $\bar{c}, \bar{c}'$ and each $y \in \{1, ..., k^* - 1\}$ expresses both z-fold belief in rationality as well $(a^p - a)$-fold belief in rationality for every order a^p.

Each type $t_2^{\bar{c}}[\bar{c}]$ and each type $t_2^{\bar{c},yz}[\bar{c}']$ in their z-th order beliefs only assign positive probability to choice-type combinations where the types are characterized as before. Hence, each type $t_2^{\bar{c}}[\bar{c}]$ and each type $t_2^{\bar{c},yz}[\bar{c}']$ only assign positive probability in their z-th order beliefs to type that express both z-fold belief in rationality and $(a^p - a)$-fold belief in rationality for every order a^p. It then follows that each type $t_2^{\bar{c}}[\bar{c}]$ and each type $t_2^{\bar{c},yz}[\bar{c}']$ then also expresses $(a^p - a + z)$-fold belief in rationality for every order a^p, $2z$-fold belief in rationality, $(a^p - a + 2z)$-fold belief in rationality for every order a^p $3z$-fold belief in rationality, and so on.

We started off with type $t_1^{c_1}[c_1]$. In the a-th order belief, the belief hierarchy induced by $t_1^{c_1}[c_1]$ exclusively assigns positive probability to the choice-type combinations $(c^a, t_2^{\bar{c}}[\bar{c}])$ where $\bar{c}$ starts with c^a. By construction choice c^a is optimal given type $t_2^{\bar{c}}[\bar{c}]$, hence type $t_1^{c_1}[c_1]$ expresses a-fold belief in rationality. Moreover, each such type $t_2^{\bar{c}}[\bar{c}]$ expresses $(a^p - a + z)$-fold belief in rationality for every order a^p, $2z$-fold belief in rationality, $(a^p - a + 2z)$-fold belief in rationality for every order a^p $3z$-fold belief in rationality, and so on. It follows that type $t_1^{c_1}[c_1]$ then also expresses $(a^p + z)$-fold belief in rationality for every order a^p, $a + z$-fold belief in rationality, $(a^p + 2z)$-fold belief in rationality for every order a^p, $a + 2z$-fold belief in rationality, and so on. Then, type $t_1^{c_1}[c_1]$ expresses on-path belief in rationality. Additionally, $t_1^{c_1}[c_1]$ was constructed such that c_1 was optimal given the type.

Hence, we have constructed an epistemic model in which c_1 is optimal given a type that expresses on-path belief in rationality.

Step 2

We will now develop a similar recursive, backward construction for scenario (iii) as we did earlier for scenarios (i) and (ii).

In Step 1 we have shown that for choice $c_1 \in C_1^{\infty}$ we can always construct a *partial* epistemic model with a type $t_1^{c_1}[c_1]$ for which c_1 is optimal and that expresses on-path belief in rationality. In Step 2 we will now do the following. We will show that if there exists a belief hierarchy expressing on-path belief in rationality for which c_1 is optimal, then there is also a belief hierarchy expressing common belief in rationality for which c_1 is optimal.

Consider a partial epistemic model $\bar{\mathcal{M}} = (T_i, b_i[t_i])_{i \in \{1,2\}}$ as constructed in Step 1. We had here for each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$ a type $t_2^{\bar{c}}[\bar{c}]$ that expressed $(a^p - a + yz)$-fold belief in rationality and $(y+1)z$-fold belief in rationality for any $y \in \mathbb{N}$. By means of a backward, recursive procedure we transform this epistemic model such that we get to a new, complete epistemic model that includes a type $t_2^m[\bar{c}, \bar{c}, 0]$ that expresses common belief in rationality and induces the same yz-th order expectation and $(a^p - a + yz)$-th order expectation for every $y \in \mathbb{N}$ as type $t_2^{\bar{c}}[\bar{c}]$ does. The recursive procedure here defines choices *and* types at the same time in each iteration.

The recursive procedure is as follows.

Iteration 0: For each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}$, define

$$d^0[\bar{c}, \bar{c}, 0] := c^a.$$

Also define for each order a^p

$$d^0[\bar{c}, \bar{c}, a^p - a] := c^{a^p}.$$

Moreover, for each $k \in \{1, 2, ..., z-1\}$ with $k \neq a^p - a$ for any order a^p, define $d^0[\bar{c}, \bar{c}, k]$ randomly:

$$d^0[\bar{c}, \bar{c}, k] := c', \text{ for } \textit{some } c' \in C_2^{\infty} \text{ if } k \text{ is even or } \textit{some } c' \in C_1^{\infty} \text{ if } k \text{ is odd.}$$

For each $y \in \{1, ..., k^* - 1\}$, define for each $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}$ with $\bar{c}' = (c^{a\prime}, ..., c^{a^x\prime})$

$$d^0[\bar{c}, \bar{c}', yz] := c^{a\prime},$$

and for each order a^p

$$d^0[\bar{c}, \bar{c}', a^p - a + yz] := c^{a^p\prime},$$

and finally for each $k \in \{1, 2, ..., z-1\}$ with $k \neq a^p - a$ for any order a^p

$$d^0[\bar{c}, \bar{c}', yz + k] := c', \text{ for } \textit{some } c' \in C_2^{\infty} \text{ if } k \text{ is even or } \textit{some } c' \in C_1^{\infty} \text{ if } k \text{ is odd.}$$

Take a sequence of types $(t_2^0[\bar{c},\bar{c},0],...,t_1^0[\bar{c},\bar{c},z-1])$ for every combination of choices $\bar{c}$. Similarly, for each $y \in \{1,...,k^*-1\}$ and each pair $\bar{c},\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ take a sequence of types $(t_2^0[\bar{c},\bar{c}',yz],...,t_1^0[\bar{c},\bar{c}',(y+1)z-1])$.

Now, for each combination of choices $\bar{c}$, define type $t_2^0[\bar{c},\bar{c},0]$ such that

$$b_2[t_2^0[\bar{c},\bar{c},0]] := (d^0[\bar{c},\bar{c},1],t_1^0[\bar{c},\bar{c},1]).$$

Then, define for each $k \in \{1,2,...,z-2\}$ type $t_i^0[\bar{c},\bar{c},k]$ with $i \in \{1,2\}$ to be such that

$$b_i[t_i^0[\bar{c},\bar{c},k]] := (d^0[\bar{c},\bar{c},k+1],t_j^0[\bar{c},\bar{c},k+1]),$$

where $j \neq i$. Finally, we define type $t_1^0[\bar{c},\bar{c},z-1]$ to be such that

$$b_1[t_1^0[\bar{c},\bar{c},z-1]](c^{a\prime},t_2^0[\bar{c},\bar{c}',z]) := b^{\bar{c}}(\bar{c}'),\ \forall \bar{c}' = (c^{a\prime},c^{a^b\prime},...,c^{a^x\prime}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

Similarly, for each pair $\bar{c},\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ and each $y \in \{1,...,k^*-1\}$, define type $t_2^0[\bar{c},\bar{c}',yz]$ to be such that

$$b_2[t_2^0[\bar{c},\bar{c}',yz]] := (d^0[\bar{c},\bar{c}',yz+1],t_1^0[\bar{c},\bar{c}',yz+1]).$$

And define for each $k \in \{1,2,...,z-2\}$ type $t_i^0[\bar{c},\bar{c}',yz+k]$ with $i \in \{1,2\}$ to be such that

$$b_i[t_i^0[\bar{c},\bar{c}',yz+k]] := (d^0[\bar{c},\bar{c}',yz+k+1],t_j^0[\bar{c},\bar{c}',yz+k+1]),$$

where $j \neq i$. Finally, we define type $t_1^0[\bar{c},\bar{c}',(y+1)z-1]$ for $y \in \{1,...,k^*-1\}$ to be such that

$$b_1[t_1^0[\bar{c},\bar{c}',(y+1)z-1]](c^{a\prime\prime},t_2^0[\bar{c},\bar{c}'',(y+1)z]) := b^{\bar{c}'}(\bar{c}''),\ \forall \bar{c}'' = (c^{a\prime\prime},...,c^{a^x\prime\prime}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

If $y = k^*-1$, define type $t_1^0[\bar{c},\bar{c}',k^*z-1]$ to be such that

$$b_1[t_1^0[\bar{c},\bar{c}',k^*z-1]](c^{a\prime\prime},t_2^0[\bar{c}'',\bar{c}'',0]) := b^{\bar{c}'}(\bar{c}''),\ \forall \bar{c}'' = (c^{a\prime\prime},...,c^{a^x\prime\prime}) \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

Note that by construction of Step 1, we have for each $y \in 1,...,k^*$ that $b^{\bar{c}'}(\bar{c}'') = b_1[t_1^0[\bar{c},\bar{c}',yz-1]](\bar{c}'',t_2^0[\bar{c},\bar{c}',yz])$ for each $\bar{c}''$. And for each order a^p we have that $b_1[t_1^0[\bar{c},\bar{c}',a^p-a-1+(y-1)z]] = (d^0[\bar{c},\bar{c}',a^p-a+(y-1)z],t_2^0[\bar{c},\bar{c}'a^p-a+(y-1)z])$. Moreover, all other types induce probability one beliefs. So for Iteration 0 we essentially take a copy of the epistemic model $\bar{\mathcal{M}}$ created in Step 1, but fill in the beliefs that were still incomplete from this step.

Iteration n $\geqslant$ 1 : For each pair $\bar{c},\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, take a sequence of types $(t_2^n[\bar{c},\bar{c}',(k^*-1)z],...,t_1^n[\bar{c},\bar{c}',k^*z-1])$. Then, define type $t_1^n[\bar{c},\bar{c}',k^*z-1]$ to be such that

$$b_1[t_1^n[\bar{c},\bar{c}',k^*z-1]](c^{a\prime\prime},t_2^{n-1}[\bar{c}'',\bar{c}'',0]) := b^{\bar{c}'}(\bar{c}''),\ \forall \bar{c}'' = (c^{a\prime\prime},...,c^{a^x\prime\prime}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}.$$

We then also define

$$d^n[\bar{c},\bar{c}',k^*z-1] := c_1',\ \text{with } c_1' \text{ optimal given the type } t_1^n[\bar{c},\bar{c}',k^*z-1].$$

We also define for each order a^p

$$d^n[\bar{c},\bar{c}',(k^*-1)z+(a^p-a)] := c^{a^p\prime}.$$

Now, for each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, define recursively for each *even* $k \in \{2, ..., z-2\}$ starting at $k = z-2$, type $t_2^n[\bar{c}, \bar{c}', (k^*-1)z + k]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}', (k^*-1)z + k]] := (d^n[\bar{c}, \bar{c}', (k^*-1)z + k + 1], t_1^n[\bar{c}, \bar{c}', (k^*-1)z + k + 1]).$$

Second, if $k \neq a^p - a$ also define

$$d^n[\bar{c}, \bar{c}', (k^*-1)z + k] := c_2^*, \text{ with } c_2^* \text{ optimal given the type } t_2^n[\bar{c}, \bar{c}', (k^*-1)z + k].$$

Third, define type $t_1^n[\bar{c}, \bar{c}', (k^*-1)z + k - 1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}', (k^*-1)z + k - 1]] := (d^n[\bar{c}, \bar{c}', (k^*-1)z + k], t_1^n[\bar{c}, \bar{c}', (k^*-1)z + k - 1].$$

Fourth, also define

$$d^n[\bar{c}, \bar{c}', (k^*-1)z + k - 1] := c_1^*, \text{ with } c_1^* \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', (k^*-1)z + k - 1]].$$

Finally, define type $t_2^n[\bar{c}, \bar{c}', (k^*-1)z]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}', (k^*-1)z]] := (d^n[\bar{c}, \bar{c}', (k^*-1)z + 1], t_1^n[\bar{c}, \bar{c}', (k^*-1)z + 1]),$$

and define

$$d^n[\bar{c}, \bar{c}', (k^*-1)z]] := c^{a\prime}.$$

Next, for each $y \in \{2, ..., k^*-2\}$, do the following iteratively, going backwards starting at $y = k^*-2$: For each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, take a sequence of types $(t_2^n[\bar{c}, \bar{c}', yz], ..., t_1^n[\bar{c}, \bar{c}', (y+1)z-1])$. Then, define type $t_1^n[\bar{c}, \bar{c}', (y+1)z - 1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}', (y+1)z - 1]](c^{a\prime\prime}, t_2^n[\bar{c}, \bar{c}'', (y+1)z]) := b^{\bar{c}'}(\bar{c}''), \ \forall \bar{c}'' = (c^{a\prime\prime}, ..., c^{a^x\prime\prime}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}.$$

We then also define

$$d^n[\bar{c}, \bar{c}', (y+1)z - 1] := c_1', \text{ with } c_1' \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', (y+1)z - 1].$$

We also define for each order a^p

$$d^n[\bar{c}, \bar{c}', yz + (a^p - a)] := c^{a^p\prime}.$$

Now, for each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, define recursively for each *even* $k \in \{2, ..., z-2\}$ starting at $k = z-2$, type $t_2^n[\bar{c}, \bar{c}', yz + k]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}', yz + k]] := (d^n[\bar{c}, \bar{c}', yz + k + 1], t_1^n[\bar{c}, \bar{c}', yz + k + 1]).$$

Second, if $k \neq a^p - a$ also define

$$d^n[\bar{c}, \bar{c}', yz + k] := c_2^*, \text{ with } c_2^* \text{ optimal given the type } t_2^n[\bar{c}, \bar{c}', yz + k].$$

Third, define type $t_1^n[\bar{c}, \bar{c}', yz + k - 1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}', yz + k - 1]] := (d^n[\bar{c}, \bar{c}', yz + k], t_1^n[\bar{c}, \bar{c}', yz + k - 1].$$

Fourth, also define

$$d^n[\bar{c}, \bar{c}', yz + k - 1] := c_1^*, \text{ with } c_1^* \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', yz + k - 1]].$$

Finally, define type $t_2^n[\bar{c}, \bar{c}', yz]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}', yz]] := (d^n[\bar{c}, \bar{c}', yz+1], t_1^n[\bar{c}, \bar{c}', yz+1]),$$

and define

$$d^n[\bar{c}, \bar{c}', yz]] := c^{a\prime}.$$

We do this iteratively for each $p \in \{0, ..., k^* - 2\}$, starting at $p = k^* - 2$.

Finally, for $y = 1$, we again do the following: For each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, take a sequence of types $(t_2^n[\bar{c}, \bar{c}, 0], ..., t_1^n[\bar{c}, \bar{c}', z-1])$. Then, define type $t_1^n[\bar{c}, \bar{c}, z-1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}, z-1]](c^{a\prime\prime}, t_2^n[\bar{c}, \bar{c}'', z]) := b^{\bar{c}}(\bar{c}''),\ \forall \bar{c}'' = (c^{a\prime\prime}, ..., c^{a^x\prime\prime}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}.$$

We then also define

$$d^n[\bar{c}, \bar{c}, z-1] := c_1',\ \text{with } c_1' \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}', z-1].$$

We also define for each order a^p

$$d^n[\bar{c}, \bar{c}, (a^p - a)] := c^{a^p}.$$

Now, for each $\bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, define recursively for each *even* $k \in \{2, ..., z-2\}$ starting at $k = z - 2$, type $t_2^n[\bar{c}, \bar{c}, k]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}, k]] := (d^n[\bar{c}, \bar{c}, k+1], t_1^n[\bar{c}, \bar{c}, k+1]).$$

Second, if $k \neq a^p - a$ also define

$$d^n[\bar{c}, \bar{c}, k] := c_2^*,\ \text{with } c_2^* \text{ optimal given the type } t_2^n[\bar{c}, \bar{c}, k].$$

Third, define type $t_1^n[\bar{c}, \bar{c}, k-1]$ to be such that

$$b_1[t_1^n[\bar{c}, \bar{c}, k-1]] := (d^n[\bar{c}, \bar{c}, k], t_1^n[\bar{c}, \bar{c}, k-1].$$

Fourth, also define

$$d^n[\bar{c}, \bar{c}, k-1] := c_1^*,\ \text{with } c_1^* \text{ optimal given the type } t_1^n[\bar{c}, \bar{c}, k-1]].$$

Finally, define type $t_2^n[\bar{c}, \bar{c}, 0]$ to be such that

$$b_2[t_2^n[\bar{c}, \bar{c}, 0]] := (d^n[\bar{c}, \bar{c}, 1], t_1^n[\bar{c}, \bar{c}, 1]),$$

and define

$$d^n[\bar{c}, \bar{c}, 0]] := c^a.$$

We have that C_1^∞ and C_2^∞ are finite sets. Additionally, z and k^* are finite orders of belief and thus k^*z is a finite order of belief as well. Finally, we have that $C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ is a finite set as well. Together, we then have that there are iterations m, n with $m > n$ such that:

$$d^m[\bar{c}, \bar{c}, k] = d^n[\bar{c}, \bar{c}, k],\ \forall \bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}, k \in \{0, 1, ..., z-1\},$$

and

$$d^m[\bar{c},\bar{c}',yz+k]=d^n[\bar{c},\bar{c}',yz+k],\ \forall \bar{c},\bar{c}'\in C_2^{a,\infty}\times ...\times C_2^{a^x,\infty}, y\in\{1,...,k^*-1\}.$$

When we find such iterations m and n, we stop the recursive procedure.

Now we create the epistemic model $\mathcal{M}^*$ from the types we have constructed in our recursive procedure. Define

$$\begin{aligned}T_2(l):=&\{t_2^l[\bar{c},\bar{c},k]:\bar{c}\in C_2^{a,\infty}\times ...\times C_2^{a^x,\infty}, k\in\{0,...,z-2\}\text{ even}\}\cup\\ &\{t_1^l[\bar{c},\bar{c}',yz+k]:\bar{c},\bar{c}'\in C_2^{a,\infty}\times ...\times C_2^{a^x,\infty},\ y\in\{1,...,k^*-1\},\ k\in\{0,...,z-2\}\text{ even}\}\end{aligned}$$

and

$$\begin{aligned}T_1(l):=&\{t_1^l[\bar{c},\bar{c},k]:\bar{c}\in C_2^{a,\infty}\times ...\times C_2^{a^x,\infty}, k\in\{0,1,...,z-1\}\text{ odd}\}\cup\\ &\{t_1^l[\bar{c},\bar{c}',yz+k]:\bar{c},\bar{c}'\in C_2^{a,\infty}\times ...\times C_2^{a^x,\infty},\ y\in\{1,...,k^*-1\},\ k\in\{0,...,z-1\}\text{ odd.}\}\end{aligned}$$

Then, let $T(l):=T_1(l)\cup T_2(l)$. Do this for every $l\in\{n,...,m\}$.

In $T(n+1)$ specifically, we re-define for each $\bar{c},\bar{c}'\in C_2^{a,\infty}\times ...\times C_2^{a^x,\infty}$ the type $t_1^{n+1}[\bar{c},\bar{c}',k^*z-1]$ to be such that

$$b_1[t_1^{n+1}[\bar{c},\bar{c}',k^*z-1]](c^{a\prime\prime},t_2^m[\bar{c}'',\bar{c}'',0]):=b^{\bar{c}'}(\bar{c}''),\forall\bar{c}''=(c^{a\prime\prime},...,c^{a^x\prime\prime})\in C_2^{a,\infty}\times ...\times C_2^{a^x,\infty}.$$

So instead of assigning positive probability to types in $T(n)$, each type $t_1^{n+1}[\bar{c},\bar{c}',k^*z-1]$ now assigns positive probability to types in $T(m)$. Then define $\mathcal{M}^*:=(\bigcup_{l\in\{n+1,...,m\}}T_i(l),b[t_i])_{i\in\{1,2\}}$.

We will show that each type in $\mathcal{M}^*$ expresses common belief in rationality. We will do so in steps.

First, for the epistemic model $\mathcal{M}^*$, we can note that for each combination of choices $\bar{c}=(c^a,c^{a^b},...,c^{a^x})$ and each $l\in\{n+1,...,m\}$, choice c^a is optimal for type $t_2^l[\bar{c},\bar{c},0]$. We can also say that for each $\bar{c}=(c^a,c^{a^b},...,c^{a^x})$, $\bar{c}'=(c^{a\prime},c^{a^b\prime},...,c^{a^x\prime})$, each $l\in\{n+1,...,m\}$ and each $y\in\{1,...,k^*-1\}$ that choice $c^{a\prime}$ is optimal for type $t_2^l[\bar{c},\bar{c}',yz]$.

Namely, from type $t_2^l[\bar{c},\bar{c},0]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_2^l[\bar{c},\bar{c},0],...,t_2^l[\bar{c},\bar{c},z-2]))$. This sequence of probability one beliefs ends at type $t_1^l[\bar{c},\bar{c},z-1]$. By our recursive backwards construction and the way we defined $b^{\bar{c}}$ in Step 1, we have that

$$marg_{C_2^\infty}b_1[t_1^l[\bar{c},\bar{c},z-1]]=b_2^{c^a}.$$

It follows then that type $t_2^l[\bar{c},\bar{c},0]$ induces a z-th order expectation that is equal to $b_2^{c^a}$. We constructed $b_2^{c^a}$ such that c^a is optimal given $b_2^{c^a}$. Hence c^a is optimal given type $t_2^l[\bar{c},\bar{c},0]$. This goes for every $l\in\{n+1,...,m\}$.

Similarly, for each $y\in\{1,...,k^*-1\}$, from type $t_2^l[\bar{c},\bar{c}',yz]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_2^l[\bar{c},\bar{c}',yz],...,t_2^l[\bar{c},\bar{c}',(y+1)z-2]))$. This sequence ends at type $t_1^l[\bar{c},\bar{c}',(y+1)z-1]$. By construction, we have that

$$marg_{C_2^\infty}b_1[t_1^l[\bar{c},\bar{c}',(y+1)z-1]]=b_2^{c^{a\prime}}.$$

It follows then that type $t_2^l[\bar{c}, \bar{c}', yz]$ induces a z-th order expectation that is equal to $b_2^{c^{a\prime}}$. We constructed $b_2^{c^{a\prime}}$ such that $c^{a\prime}$ is optimal given $b_2^{c^{a\prime}}$. Hence $c^{a\prime}$ is optimal given type $t_2^l[\bar{c}, \bar{c}', yz]$. This goes for every $l \in \{n+1, ..., m\}$.

Second, we can also note that for each combination of choices $\bar{c}$, each order a^p and each $l \in \{n+1, ..., m\}$ we have that choice c^{a^p} is optimal for type $t_2^l[\bar{c}, \bar{c}, a^p - a]$. Moreover, for each pair $\bar{c}, \bar{c}'$, each order a^p, each $l \in \{n+1, ..., m\}$ and each $y \in \{1, ..., k^* - 1\}$ we have that choice $c^{a^p\prime}$ is optimal for the type $t_2^l[\bar{c}, \bar{c}', a^p - a + yz]$.

Namely, from type $t_2^l[\bar{c}, \bar{c}, 0]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_2^l[\bar{c}, \bar{c}, 0], ..., t_2^l[\bar{c}, \bar{c}, z-2])$. This sequence of probability one beliefs ends at type $t_1^l[\bar{c}, \bar{c}, z-1]$. By our recursive backwards construction, we have that

$$b_1[t_1^l[\bar{c}, \bar{c}, z-1](d^l[\bar{c}, \bar{c}', z], t_2^{l-1}[\bar{c}, \bar{c}', z]) = b^{\bar{c}}(\bar{c}'), \forall \bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime})' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

By Step 1 we defined $b^{\bar{c}}$ as the joint probability distribution of the z-th order expectations $b_2^{c^{a^p}}$ for all orders a^p. Now, from each type $t_2^{l-1}[\bar{c}, \bar{c}', 0]$ there again follows a sequence of probability one beliefs up to at least type $t_2^{l-1}[\bar{c}, \bar{c}', a^p - a - 1 + z]$. And type $t_2^{l-1}[\bar{c}, \bar{c}', a^p - a - 1 + z]$ assigns by construction of our recursive procedure probability one to choice $c^{a^p\prime}$. Taken together, it follows that type $t_2^l[\bar{c}, \bar{c}, a^p - a]$ induces a z-th order expectation that is equal to $b_2^{c^{a^p}}$. By construction of Step 1, we have that choice c^{a^p} is optimal given $b_2^{c^{a^p}}$. Hence, choice c^{a^p} is also optimal given type $t_2^l[\bar{c}, \bar{c}, a^p - a]$.

Similarly, for each $y \in \{1, ..., k^* - 1\}$, from type $t_2^l[\bar{c}, \bar{c}', yz]$ there follows a sequence of probability one beliefs, induced by the sequence of types $(t_2^l[\bar{c}, \bar{c}', yz], ..., t_2^l[\bar{c}, \bar{c}', (y+1)z - 2])$. This sequence of probability one beliefs ends at type $t_1^l[\bar{c}, \bar{c}', (y+1)z - 1]$. By our recursive backwards construction, we have that

$$b_1[t_1^l[\bar{c}, \bar{c}', (y+1)z - 1](d^l[\bar{c}, \bar{c}'', (y+1)z], t_2^{l-1}[\bar{c}, \bar{c}'', (y+1)z]) = b^{\bar{c}'}(\bar{c}''),$$
$$\forall \bar{c}'' = (c^{a\prime\prime}, c^{a^b\prime\prime}, ..., c^{a^x\prime\prime})' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}.$$

By Step 1 we defined $b^{\bar{c}'}$ as the joint probability distribution of the z-th order expectations $b_2^{c^{a^p\prime}}$ for all orders a^p. Now, from each type $t_2^{l-1}[\bar{c}, \bar{c}'', (y+1)z]$ there again follows a sequence of probability one beliefs up to at least type $t_2^{l-1}[\bar{c}, \bar{c}'', a^p - a - 1 + (y+1)z]$. And type $t_2^{l-1}[\bar{c}, \bar{c}'', a^p - a - 1 + (y+1)z]$ assigns by construction of our recursive procedure probability one to choice $c^{a^p\prime\prime}$. Taken together, it follows that type $t_2^l[\bar{c}, \bar{c}', a^p - a + yz]$ induces a z-th order expectation that is equal to $b_2^{c^{a^p\prime}}$. By construction of Step 1, we have that choice $c^{a^p\prime}$ is optimal given $b_2^{c^{a^p\prime}}$. Hence, choice $c^{a^p\prime}$ is also optimal given type $t_2^l[\bar{c}, \bar{c}', a^p - a + yz]$.

Third, we can also show the following is true.

Claim 4. *Consider the epistemic model $\mathcal{M}^*$. For each $l \in \{n+1, ..., m\}$, each $k \in \{1, 2, ..., z-1\}$ for $k \neq a^p - a$ for any order a^p and each combination of choices $\bar{c} \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}$, each choice $d^l[\bar{c}, \bar{c}, k]$ is optimal given the type $t_i^l[\bar{c}, \bar{c}, k]$ with $i \in \{1, 2\}$. Moreover, for each $y \in \{1, ..., k^* - 1\}$, for each $l \in \{n+1, ..., m\}$, each $k \in \{1, 2, ..., z-1\}$ for $k \neq a^p - a$ for any order a^p and each pair $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times C_2^{a^b,\infty} \times ... \times C_2^{a^x,\infty}$, each choice $d^l[\bar{c}, \bar{c}', yz + k]$ is optimal given the type $t_i^l[\bar{c}, \bar{c}', yz + k]$ with $i \in \{1, 2\}$.*

Proof of claim. We start of with the epistemic model we created when ending the recursive procedure, but *before* $\mathcal{M}^*$ was created.

For each $k \in \{0, 1, ..., z-2\}$ and each $\bar{c}'$ we have by construction that

$$b_i[t_i^n[\bar{c}', \bar{c}', k](d^n[\bar{c}', \bar{c}', k+1], t_j^n[\bar{c}', \bar{c}', k+1]) = 1 = b_i[t_i^m[\bar{c}', \bar{c}', k](d^m[\bar{c}', \bar{c}', k+1], t_j^m[\bar{c}', \bar{c}', k+1]),$$

with $d^n[\bar{c}', \bar{c}', k+1] = d^m[\bar{c}', \bar{c}', k+1]$. Note that these were the n and m that determined when to stop our recursive procedure. Moreover, for each $k \in \{0, 1, ..., a-2\}$, each $\bar{c}', \bar{c}^* \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ and each $y \in \{1, ..., k^* - 1\}$ we also have by construction

$$b_i[t_i^n[\bar{c}', \bar{c}^*, yz+k](d^n[\bar{c}', \bar{c}^*, yz+k+1], t_j^n[\bar{c}', \bar{c}^*, yz+k+1]) = 1 =$$
$$b_i[t_i^m[\bar{c}', \bar{c}^*, yz+k](d^m[\bar{c}', \bar{c}^*, yz+k+1], t_j^m[\bar{c}', \bar{c}^*, yz+k+1]),$$

with $d^n[\bar{c}', \bar{c}^*, yz+k+1] = d^m[\bar{c}', \bar{c}^*, yz+k+1]$. Additionally, we have by construction that

$$b_1[t_1^n[\bar{c}', \bar{c}', z-1](d^n[\bar{c}', \bar{c}'', z], t_2^n[\bar{c}', \bar{c}'', z]) = b^{\bar{c}'}[\bar{c}''] = b_1[t_1^m[\bar{c}', \bar{c}', z-1](d^m[\bar{c}', \bar{c}'', z], t_2^m[\bar{c}', \bar{c}'', z]),$$

for each $\bar{c}'' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$. For each $y \in \{1, ..., k^* - 2\}$ we also have that

$$b_1[t_1^n[\bar{c}', \bar{c}^*, yz-1](d^n[\bar{c}', \bar{c}'', yz], t_2^n[\bar{c}', \bar{c}'', yz]) = b^{\bar{c}^*}[\bar{c}''] =$$
$$b_1[t_1^m[\bar{c}', \bar{c}^*, yz-1](d^m[\bar{c}', \bar{c}'', yz], t_2^m[\bar{c}', \bar{c}'', yz]),$$

for each $\bar{c}'' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$. Finally, we have that

$$b_1[t_1^n[\bar{c}', \bar{c}^*, k^*z-1](d^{n-1}[\bar{c}'', c_1'', 0], t_2^{n-1}[\bar{c}'', \bar{c}'', 0]) = b^{\bar{c}^*}[\bar{c}''] =$$
$$b_1[t_1^m[\bar{c}', \bar{c}^*, k^*z-1](d^{m-1}[\bar{c}'', \bar{c}'', 0], t_2^{m-1}[\bar{c}'', \bar{c}'', 0]),$$

for each $\bar{c}'' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$.

Then, for each $\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, the pair of types $t_2^m[\bar{c}', \bar{c}', 0]$ and $t_2^n[\bar{c}', \bar{c}', 0]$ induce the same k^*z-th order *belief*. To see why this is the case, we can employ a recursive argument, for each $y \in \{1, ..., k^* - 1\}$ starting at $y = k^* - 1$.

We can first note that the pair of types $t_2^m[\bar{c}', \bar{c}^*, (k^*-1)z]$ and $t_2^n[\bar{c}', \bar{c}^*, (k^*-1)z]$ for each $\bar{c}', \bar{c}^*$ induce the same z-th order belief. Namely, from the beginning of the proof of this claim we know that types $t_i^m[\bar{c}', \bar{c}^*, (k^*-1)z+k]$ and $t_i^n[\bar{c}', \bar{c}^*, (k^*-1)z+k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., z-2\}$ induce a probability one belief. Moreover, the first-order belief induced by type $t_i^m[\bar{c}', \bar{c}^*, (k^*-1)z+k]$ for each $k \in \{1, ..., z-1\}$ is equal to the first-order belief induced by type $t_i^n[\bar{c}', \bar{c}^*, (k^*-1)z+k]$. As a result, types $t_i^m[\bar{c}', \bar{c}^*, (k^*-1)z+k]$ and $t_i^n[\bar{c}', \bar{c}^*, (k^*-1)z+k]$ induce the same z-th order *belief*.

Now recall, for each $\bar{c}', \bar{c}^* \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, we have that

$$b_1[t_1^n[\bar{c}', \bar{c}^*, (k^*-1)a-1](d^n[\bar{c}', \bar{c}'', (k^*-1)z], t_2^n[\bar{c}', \bar{c}'', (k^*-1)z]) = b_1^{\bar{c}^*}[\bar{c}''] =$$
$$b_1[t_1^m[\bar{c}', \bar{c}^*, (k^*-1)z-1](d^m[\bar{c}', \bar{c}'', (k^*-1)z], t_2^m[\bar{c}', \bar{c}'', (k^*-1)z]).$$

Both types $t_1^m[\bar{c}', \bar{c}^*, (k^*-1)z-1]$ and $t_1^n[\bar{c}', \bar{c}^*, (k^*-1)z-1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same z-th order belief. Hence, both types induce the same $(z+1)$-th order belief.

Now we can employ our recursive argument, starting at $y = k^* - 2$. For $y \in \{1, ..., k^* - 2\}$, assume that types $t_1^m[\bar{c}', \bar{c}^*, (y+1)z-1]$ and $t_1^n[\bar{c}', \bar{c}^*, (y+1)z-1]$ induce the same $((k^*-y-1)z+1)$-th order belief. Then types $t_2^m[\bar{c}', \bar{c}^*, yz]$ and $t_2^n[\bar{c}', \bar{c}^*, yz]$ induce the same $(k^*-p)a$-th order

belief. Namely, from the beginning of the proof of this claim we have that types $t_i^m[\bar{c}', \bar{c}^*, yz + k]$ and $t_i^n[\bar{c}', \bar{c}^*, yz + k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., z - 2\}$ induce a probability one belief and moreover induce the same first-order belief. Therefore, they induce the same $(z-1)$-th order belief. Additionally, types $t_2^m[\bar{c}', \bar{c}^*, (y + 1)z - 2]$ and $t_2^n[\bar{c}', \bar{c}^*, (y + 1)z - 2]$ assign probability one to types that by assumption induce the same $((k^* - y)z + 1)$-th order belief. It follows then that types $t_2^m[\bar{c}', \bar{c}^*, yz]$ and $t_2^n[\bar{c}', \bar{c}^*, yz]$ induce the same $(k^* - y)z$-th order belief.

Now recall that for each $\bar{c}', \bar{c}^* \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$, we have that

$$b_1[t_1^n[\bar{c}', \bar{c}^*, pa - 1]](d^n[\bar{c}', \bar{c}'', yz], t_2^n[\bar{c}', \bar{c}'', yz]) = b_1^{\bar{c}_1}[c_1''] =$$
$$b_1[t_1^m[\bar{c}', \bar{c}^*, pa - 1]](d^m[\bar{c}', \bar{c}'', yz], t_2^m[\bar{c}', \bar{c}'', yz])$$

Both types $t_1^m[\bar{c}', \bar{c}^*, pa - 1]$ and $t_1^n[\bar{c}', \bar{c}^*, pa - 1]$ thus assign exactly the same probability to choice-type combinations where the choice is equal and the type induces the same $(k^* - y)z$-th order belief. Hence, both types induce the same $((k^* - y)z + 1)$-th order belief.

Following the same argument, we can establish that types $t_2^m[\bar{c}', \bar{c}', 0]$ and $t_2^n[\bar{c}', \bar{c}', 0]$ induce the same k^*z-th order belief. From the above we know that types $t_1^m[\bar{c}', \bar{c}', z - 1]$ and $t_1^n[\bar{c}', \bar{c}', z - 1]$ induce the same $((k^* - 1)y + 1)$-th order belief. From the beginning of the proof of this claim we have that types $t_i^m[\bar{c}', \bar{c}', k]$ and $t_i^n[\bar{c}', \bar{c}', k]$ with $i \in \{1, 2\}$ for each $k \in \{1, ..., z - 2\}$ induce a probability one belief and moreover induce the same first-order belief. Therefore, they induce the same $(z-1)$-th order belief. Additionally, types $t_2^m[\bar{c}', \bar{c}', z-2]$ and $t_2^n[\bar{c}', \bar{c}', z-2]$ assign probability one to types that by the above recursive argument induce the same $((k^* - 1)z + 1)$ order belief. It follows then that types $t_2^m[\bar{c}', \bar{c}', 0]$ and $t_2^n[\bar{c}', \bar{c}', 0]$ induce the same k^*z-th order belief. This goes for each $\bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$.

Denote type $t_1^{n+1}[\bar{c}, \bar{c}^*, k^*z - 1]$ that results from our recursive backwards procedure but *before* constructing $\mathcal{M}^*$ by $\bar{t}_1^{n+1}[\bar{c}, \bar{c}^*, k^*z - 1]$. In contrast, let the same type that does result from constructing $\mathcal{M}^*$ still be denoted as $t_1^{n+1}[\bar{c}, \bar{c}^*, k^*z - 1]$. Now, we have for each $\bar{c} \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$

$$b_1[\bar{t}_1^{n+1}[\bar{c}, \bar{c}^*, k^*z - 1]](c^{a\prime}, t_2^n[\bar{c}', \bar{c}', 0]) = b_1[t_1^{n+1}[\bar{c}, \bar{c}^*, k^*z - 1]](c^{a\prime}, t_1^n[\bar{c}', \bar{c}', 0]),$$
$$\forall \bar{c}' = (c^{a\prime}, ..., c^{x\prime}) \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}.$$

It thus follows that each such type $t_1^{n+1}[\bar{c}, \bar{c}^*, k^*z - 1]$ induces the same $(k^*z + 1)$-th order belief in $\mathcal{M}^*$ as it did before $\mathcal{M}^*$ was constructed. All the remaining types in $\bigcup_{l \in \{n+1,...,m\}} T(l)$ remained unchanged when $\mathcal{M}^*$ was constructed: they induce exactly the same belief over choice-type combinations as before. As a result, all types in $\bigcup_{l \in \{n+1,...,m\}} T(l)$ induce at least the same $(k^*z + 1)$-th order belief in $\mathcal{M}^*$ as before $\mathcal{M}^*$ was constructed.

In our backward construction procedure of types and choices, before creating $\mathcal{M}^*$, we constructed each $d^l[\bar{c}, \bar{c}', k]$ for each $l \in \{n + 1, ..., m\}$, $k \in \{1, ..., k^*z - 1\}$ and $\bar{c}, \bar{c}' \in C_2^{a,\infty} \times ... \times C_2^{a^x,\infty}$ such that it is optimal given type $t_i^l[\bar{c}, \bar{c}', k]$. Now, we have that the maximum directly utility-relevant order of belief for any player is k^* and that each type $t_i^l[\bar{c}, \bar{c}', k]$ at least induces exactly the same $(k^*z + 1)$-th order belief in $\mathcal{M}^*$ as it did before constructing $\mathcal{M}^*$. Hence, we also have in $\mathcal{M}^*$ that $d^l[\bar{c}, \bar{c}', k]$ is optimal given $t_i^l[\bar{c}, \bar{c}', k]$. This completes the proof of this claim.

Since each type in $\mathcal{M}^*$ only assigns positive probability to choice-type combinations of the likes of $(d^l[\bar{c}, \bar{c}', k], t_i^l[\bar{c}, \bar{c}', k]])$ for $k \in \{0, 1, ..., k^*a - 1\}$, each type only assigns positive probability to choice-type combinations where the choice is optimal given the type. Hence each type in $\mathcal{M}^*$

expresses 1-fold belief in rationality. Therefore also each type in $\mathcal{M}^*$ expresses common belief in rationality.

We have now the following result. For each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$, we have for choice c^a constructed type $t_2^m[\bar{c}, \bar{c}, 0]$ such that c^a is optimal given that type. This type also expresses common belief in rationality in $\mathcal{M}^*$ Additionally, for each pair $\bar{c}, \bar{c}'$ with $\bar{c}' = (c^{a\prime}, c^{a^b\prime}, ..., c^{a^x\prime})$, the type $t_1^m[\bar{c}, \bar{c}', a^p - a - 1 + yz]$ always assigns probability one to choice $c^{a^p\prime}$.

As a final step, we extend this finite epistemic model in the following way. Consider again the type $t_1^{c_1}[c_1]$ we fixed in Step 1 of this proof. Choice c_1 is optimal given some higher-order expectation $b_1^{c_1} \in \Delta(C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty})$.

We have from Step 1 that $t_1^{c_1}[c_1]$ is at the start of the following sequence of types: $(t_1^{c_1}[c_1], ..., t_1^{c_1,a-1}[c_1])$. Type $t_1^{c_1}[c_1]$ is such that it assigns probability one to type $t_2^{c_1,1}[c_1]$ and for each $n \in \{1, 2, ..., a-2\}$ we have that type $t_i^{c_1,n}[c_1]$ is such that it assigns probability one to type $t_j^{c_1,n+1}[c_1]$.

Second, recall that each combination of choices $\bar{c} = (c^a, c^{a^b}, ..., c^{a^x})$ was derived from a different combination of choices in $C_2^{a,\infty} \times C_2^{b,\infty} \times ... \times C_2^{x,\infty}$. Then, by construction of Step 1 we had for each $(c^a, c^b, ..., c^x)$ and each combination of choices $\bar{c}$ that is derived from it, that type $t_1^{c_1,a-1}[c_1]$ is such that

$$b_1[t_1^{c_1,a-1}[c_1]](c^a, t_2^m[\bar{c}', \bar{c}', 0]) := \begin{cases} b_1^{c_1}(c^a, c^b, ..., c^x), \text{ if } \bar{c}' = \bar{c}, \\ 0, \text{ otherwise,} \end{cases}$$

where type $t_2^m[\bar{c}', \bar{c}', 0]$ now replaces type $t_2^{\bar{c}}[\bar{c}']$ from Step 1. Taken together, we have that type $t_1^{c_1,a-1}[c_1]$ is constructed such that choice c_1 is optimal given type $t_1^{c_1}[c_1]$. Moreover, type $t_1^{c_1,a-1}[c_1]$, by construction of $\mathcal{M}^*$ only assigns positive probability to choice-type combinations where the choice is optimal given the type and the type expresses common belief in rationality.

Finally, for each $n \in \{1, 2, ..., a-2\}$, do the following in a stepwise manner, starting at $n = a-2$. Take type $t_i^{c_1,n}[c_1]$. Let $b_i[t_i^{c_1,n}[c_1]] := (c', t_j^{c_1,n+1}[c_1])$, with c' being optimal given type $t_i^{c_1,n+1}[c_1]$. Likewise, let $b_1[t_1^{c_1}[c_1] := (c', t_2^{c_1,1}[c_1])$, with c' such that it is optimal given $t_2^{c_1,1}[c_1]$. Then, type $t_i^{c_1,n}[c_1]$ for each n and type $t_1^{c_1}[c_1]$ express belief in the opponent's rationality. As such, we have iteratively connected type $t_1^{c_1}[c_1]$ exclusively to types that express common belief in rationality. Hence, type $t_1^{c_1}[c_1]$ expresses common belief in rationality. Call the resulting epistemic model $\mathcal{M}^{**}$.

Thus, we constructed a finite epistemic model with a type that expresses common belief in rationality for which choice c_1 is optimal.

In Step 1 we have shown that for every choice $c_1 \in C_1^\infty$ we can construct a partial epistemic model with a type that expresses on-path belief in rationality and that is such that choice c_1 is optimal. In Step 2 we showed that we are then also able to construct a finite, epistemic model with a type that expresses common belief in rationality and that is such that choice c_1 is optimal. This concludes the proof for Scenario (*iii*). This also concludes the proof for Lemma 4 as a whole.

□

References

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

Attanasi, G., Battigalli, P., and Nagel, R. (2013). Disclosure of belief-dependent preferences in the trust game. *IGIER Working Paper no. 506*.

Bach, C. and Perea, A. (2016). Incomplete information and generalized iterative strict dominance. *Epicenter Working Paper No. 7*.

Battigalli, P., Charness, G., and Dufwenberg, M. (2013). Deception: the role of guilt. *Journal of Economic Behaviour and Organization*, 93, 227–232.

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

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

Battigalli, P., Dufwenberg, M., and Smith, A. (2015). Frustration and anger in games. *IGIER Working Paper no. 539*.

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

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, 52(4), 1007–1028.

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

Brandenburger, A. and Dekel, E. (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*.

Caplin, A. and Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*, 116, 55–79.

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

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

Charness, G., Naef, M., and Sontuoso, A. (2019). Opportunistic conformism. *Journal of Economic Theory*, 180, 100–134.

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

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behaviour*, 47(2), 269–298.

Dufwenberg, Jr., M. and Dufwenberg, M. (2018). Lies in disguise - a theoretical analysis of cheating. *Journal of Economic Theory*, 175.

Ely, J., Frankel, A., and Kamenica, E. (2015). Suspense and surprise. *Journal of Political Economy*, 123, 215–260.

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

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

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of a lie. *American Economic Review*, 108(2), 419–453.

Jagau, S. and Perea, A. (2017). Common belief in rationality in psychological games. *Epicenter Working Paper No. 10*.

Jagau, S. and Perea, A. (2018). Expectation-based psychological games and psychological expected utility. Work in progress.

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

Li, J. (2008). The power of conventions: a theory of social preferences. *Journal of Economic Behaviour and Organizaiton*, 65(3), 489–505.

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

Perea, A. (2014). Belief in the opponents' future rationality. *Games and Economic Behaviour*, 83, 231–254.

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

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

Spohn, W. (1982). How to make sense of game theory. In Stegmuller, W., Balzer, W., and Spohn, W., editors, *Philosophy of economics*, pages 239–270. Heidelberg and New York: Springer Verlag.

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