# Chapter 8

# Common Belief in Rationality in Psychological Games

In the previous chapters, the preference relation over your choices only depended on your belief about the opponents' choices – that is, your first-order belief – but not on your higher-order beliefs. There are real-life situations, however, where the preferences over your choices depend on your higher-order beliefs, for instance, on what you believe that the opponent believes about your choice.

As an example, suppose that you wish to *surprise* the opponent by the choice you make. Then, you wish to make a choice such that you believe that the opponent believes that you make a different choice than the one you are considering. As a consequence, the utility you derive from a making a choice depends on what you believe that the opponent believes about the choice you are going to make, which is your second-order belief.

Or consider a scenario where you wish to *meet the opponent's expectations*. Then, your objective is to make a choice which is "at least as good", from the opponent's perspective, as the choice he expects you to make. Also in this case, the utility you derive from making a certain choice depends on what you believe that the opponent believes about the choice you are going to make.

Such situations, where the preferences over your own choices depend on second-order or even higher-order beliefs, are called *psychological games.* We will see that these games can still be represented by collections of one-person decision problems, but the *states* in these one-person decision problems must contain more than just the opponents' choice-combinations. The reason is the following: Since the preferences of a player may depend on second-order or even higher-order beliefs, the relevant uncertainty not only concerns what the other players *do*, but also what the other players *believe*. If, for instance, your utility depends on your second-order belief, then the relevant uncertainty for you consists of the opponent's choice, but also of the opponent's belief about your choice. Therefore, the opponent's belief about your choice must be part of the states in your decision problem.

As will become clear, the idea of common belief in rationality in psychological games can be formalized in almost the same way as for standard games in Chapter 3. In fact, the only difference lies in the definition of an optimal choice for a type. Also belief hierarchies, beliefs diagrams and epistemic

| **You** | *blue* | *green* | *red* |
|---|---|---|---|
| *blue* | 0 | 3 | 3 |
| *green* | 2 | 0 | 2 |
| *red* | 1 | 1 | 0 |

| **Barbara** | *blue* | *green* | *red* |
|---|---|---|---|
| *blue* | 0 | 2 | 2 |
| *green* | 1 | 0 | 1 |
| *red* | 3 | 3 | 0 |

Table 8.1.1 Baseline decision problems for "Surprising Barbara"

models look exactly the same as for standard games.

What will be fundamentally different from previous chapters is the recursive elimination procedure that characterizes the choices that can rationally be made under common belief in rationality. Whereas previous procedures recursively eliminated choices and states from the different decision problems, the procedure from this chapter is more complicated: In every round it eliminates *combinations* of *choices* and (possibly higher-order) *beliefs.*

In Chapter 8 of the online appendix we will discuss some economic applications of the theory in this chapter.

# 8.1 Example

As announced above, we will focus on situations where the preferences over your own choices depend on your second-order, or even higher-order, beliefs. We start with an example that represents such a situation.

**Example 8.1: Surprising Barbara.**

Like in Example 3.1, "Going to a party", you are invited to a party, and so is Barbara. The problem is, again, which color to wear. Suppose that you and Barbara can choose between the colors *blue, green* and *red.* You prefer *blue* to *green,* and *green* to *red,* whereas Barbara prefers *red to blue,* and *blue* to *green.* However, similarly to Example 3.1, you both dislike it when the other person wears the same color. Assume that your and Barbara's conditional preference relation are represented by the two decision problems in Table 8.1.1. Which colors would you rationally wear under common belief in rationality?

Note that for you, the color *red* is strictly dominated by the randomized choice $(0.4)\cdot$ *blue* $+ (0.6)\cdot$ *green,* and hence we know from Theorem 2.6.1 that wearing *red* cannot be optimal for you for any belief. Hence, if Barbara believes in your rationality, then she believes you will not choose *red.* But then, Barbara would definitely choose *red,* as *red* is her favorite color. Summarizing, we thus see that if you believe in Barbara's rationality, and believe that Barbara believes in your rationality, then you expect Barbara to choose *red.* As a consequence, you would choose *blue.* Hence, under common belief in rationality, your only rational choice is to wear *blue.*

However, after a few parties you have the feeling that you have become very predictable, by always wearing *blue,* and this annoys you. From now on, you do not only want to wear a different color than Barbara, but you also wish to *surprise* Barbara by the color that you wear. We model this surprise component as follows: If you consider wearing *blue,* and you believe, with probability 1, that Barbara believes, with probability 1, that you will wear a different color than *blue,* then the *surprise utility* from wearing *blue* would be 3, as specified in Table 8.1.2 in the first row for you. If, on the other

| **You** | Barbara believes you choose *blue* | *green* | *red* |
|---|---|---|---|
| *blue* | 0 | 3 | 3 |
| *green* | 2 | 0 | 2 |
| *red* | 1 | 1 | 0 |

| **Barbara** | you believe Barbara chooses *blue* | *green* | *red* |
|---|---|---|---|
| *blue* | 0 | 2 | 2 |
| *green* | 1 | 0 | 1 |
| *red* | 3 | 3 | 0 |

Table 8.1.2 Surprise utilities for "Surprising Barbara"

| **You** | $(b,b)$ | $(b,g)$ | $(b,r)$ | $(g,b)$ | $(g,g)$ | $(g,r)$ | $(r,b)$ | $(r,g)$ | $(r,r)$ |
|---|---|---|---|---|---|---|---|---|---|
| *blue* | 0 | 3 | 3 | 3 | 6 | 6 | 3 | 6 | 6 |
| *green* | 4 | 2 | 4 | 2 | 0 | 2 | 4 | 2 | 4 |
| *red* | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 0 |

| **Barbara** | $(b,b)$ | $(b,g)$ | $(b,r)$ | $(g,b)$ | $(g,g)$ | $(g,r)$ | $(r,b)$ | $(r,g)$ | $(r,r)$ |
|---|---|---|---|---|---|---|---|---|---|
| *blue* | 0 | 2 | 2 | 2 | 4 | 4 | 2 | 4 | 4 |
| *green* | 2 | 1 | 2 | 1 | 0 | 1 | 2 | 1 | 2 |
| *red* | 6 | 6 | 3 | 6 | 6 | 3 | 3 | 3 | 0 |

Table 8.1.3 Decision problems for "Surprising Barbara"

hand, you believe, with probability 1, that Barbara believes, with probability 1, that you will wear *blue,* then the *surprise utility* by wearing *blue* would be 0, because you believe it to be no surprise at all when you dress in *blue.* In a similar way we can derive the *surprise utility* for the other two colors, which leads to the first matrix in Table 8.1.2. Note that the surprise utility depends on your *second-order* belief, that is, what you believe that Barbara believes about your color choice.

Suppose now that your utility from wearing a certain color equals the sum of the *baseline utility* from Table 8.1.1 and the *surprise utility* from Table 8.1.2. This reflects the fact that you care equally much about wearing a *different color* than Barbara on the one hand, and *surprising* Barbara on the other hand. Then, your utility of wearing a given color corresponds to the first matrix in Table 8.1.3. Here, the state $(b,g)$ represents the situation where Barbara chooses *blue,* and Barbara believes, with probability 1, that you choose *green.* If you choose *blue* in that case, then your baseline utility would be 0, because you are wearing the same color as Barbara, whereas your surprise utility is 3, since you believe that Barbara believes that you will wear *green* and not *blue.* In the same way we can interpret the other states, and derive the utilities at the other entries in the matrix. Note that a state consists of a choice and a first-order belief of Barbara, since your preferences over your own colors depend both on Barbara's choice (because of the baseline utilities) and on Barbara's first-order belief (because of the surprise utilities).

We assume that Barbara's conditional preference relation is similar to yours, in the sense that also her utility is the sum of her baseline utility and her surprise utility. That is, also Barbara would like to surprise you by the choice of her color. Naturally, Barbara's surprise utilities are given by the second matrix in Table 8.1.2. This, in turn, leads to Barbara's decision problem given by the second matrix in Table 8.1.3. Please verify this.

Let us have a closer look at your decision problem in Table 8.1.3. The numbers describe the utilities you would obtain for every state $(c_2, c_1)$, which represent situations where you believe, with probability 1, that Barbara chooses the color $c_2$, and you believe, with probability 1, that Barbara

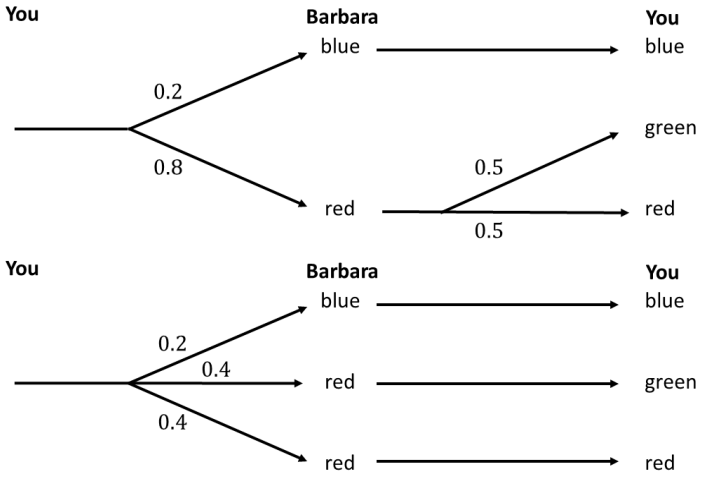



Figure 8.1.1 Probabilistic first- and second-order belief for "Surprising Barbara"

believes, with probability 1, that you choose the color $c_1$. But what would your expected utilities be if you hold probabilistic first- and second-order beliefs?

Let us consider an example first. Suppose you consider choosing *green,* and your first- and second-order beliefs are given by the partial beliefs diagram in the upper half of Figure 8.1.1. That is, you assign probability 0.2 to the event that Barbara wears *blue* and believes, with probability 1, that you also wear *blue,* and you assign probability 0.8 to the event that Barbara wears *red* and assigns probability 0.5 to you wearing *green* and *red.* What would your expected utility of wearing *green* be?

In expectation, your first- and second-order belief assign probability 0.2 to the event that Barbara wears *blue* and believes that you wear *blue.* In that sense, it assigns probability 0.2 to the state $(b, b)$. Moreover, in expectation they assign probability $(0.8) \cdot (0.5) = 0.4$ to the event that Barbara wears *red* and believes that you wear *green.* That is, they assign probability 0.4 to the state $(r, g)$. Finally, the first- and second-order belief assign, in expectation, probability $(0.8) \cdot (0.5) = 0.4$ to the event that Barbara wears *red* and believes that you wear *red.* Hence, they assign probability 0.4 to the state $(r, r)$. Summarizing, we see that your first- and second-order belief in Figure 8.1.1 induce the probability distribution

$$(0.2) \cdot (b, b) + (0.4) \cdot (r, g) + (0.4) \cdot (r, r)$$

over the states. This probability distribution is called the *second-order expectation* induced by the first- and second-order belief in Figure 8.1.1. Given this second-order expectation and the utilities from Table 8.1.3, the expected utility from wearing *green* is thus

$$(0.2) \cdot 4 + (0.4) \cdot 2 + (0.4) \cdot 4 = 3.2.$$

**Question 8.1.1** *Consider the probabilistic first- and second-order belief in the upper half of Figure 8.1.1. What are the expected utilities from choosing blue and red under this belief? What choice is optimal for you under that belief?*

Consider next the probabilistic first- and second-order belief from the lower half of Figure 8.1.1. This belief is different from the one above, since now you assign probability 0.2 to the event that Barbara wears *blue* and believes, with probability 1, that you also wear *blue,* you assign probability 0.4 to the event that Barbara wears *red* and believes, with probability 1, that you wear *green,* and you

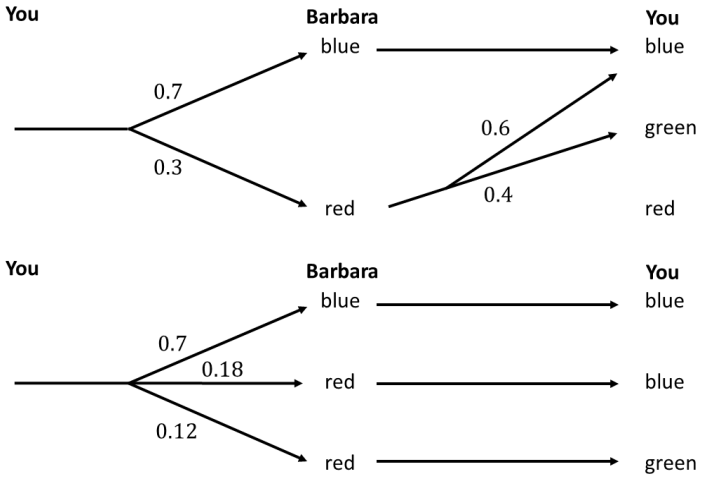



Figure 8.1.2 Beliefs for "Surprising Barbara" in Question 8.1.2

assign probability 0.4 to the event that Barbara wears *red* and believes, with probability 1, that you also wear *red.* In particular, in this belief you only consider events where Barbara assigns probability 1 to one particular choice of yours. In other words, you are convinced that Barbara is certain about your choice of color. This is not true for your belief in the upper half of Figure 8.1.1. Indeed, in that belief you assign probability 0.8 to the event that Barbara assigns probability 0.5 to you wearing *green* and *red.* That is, you believe, with probability 0.8, that Barbara is truly uncertain about your choice of color.

Nevertheless, we will see that both beliefs induce the same second-order expectation, and hence will yield the same expected utilities for each of your colors. To see this, consider the belief from the lower half of Figure 8.1.1. As we have seen above, it assigns probability 0.2 to the event that Barbara wears *blue* and believes, with probability 1, that you also wear *blue,* it assigns probability 0.4 to the event that Barbara wears *red* and believes, with probability 1, that you wear *green,* and it assigns probability 0.4 to the event that Barbara wears *red* and believes, with probability 1, that you also wear *red.* As such, it assigns, in expectation, probability 0.2 to the state $(b, b)$, probability 0.4 to the state $(r, g)$ and probability 0.4 to the state $(r, r)$. Hence, the induced second-order expectation is

$$(0.2) \cdot (b, b) + (0.4) \cdot (r, g) + (0.4) \cdot (r, r),$$

the same as for the belief in the upper half of Figure 8.1.1.

We thus see that the two beliefs, although different, induce the same second-order expectation, and thereby the same expected utility for each of your choices. As a consequence, the preference relation over your choices will be the same, no matter whether you hold the belief in the upper part or the belief in the lower part of Figure 8.1.1. The reason is that your expected utility only depends on the second-order expectation, which is a probability distribution over the states, and not on the full specification of the first- and second-order belief.

**Question 8.1.2** *Consider the first- and second-order belief from the upper half of Figure 8.1.2. What is the second-order expectation induced by this belief? Show that the belief in the lower part of the figure is different, but induces the same second-order expectation. Find, for every choice of yours, the expected utility it induces under either of these beliefs. What is your optimal choice?*

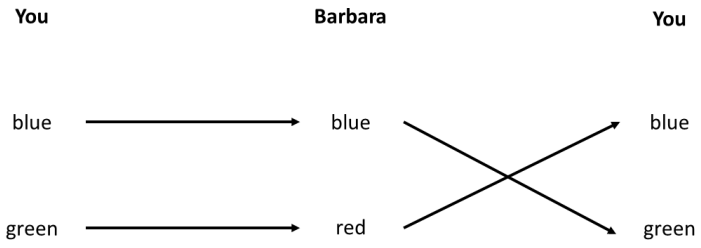



Figure 8.1.3 Beliefs diagram for "Surprising Barbara"

Now that we have represented the situation at hand by a collection of one-person decision problems – one for you and one for Barbara – we ask: Which color(s) can you rationally choose under common belief in rationality?

Note that in your decision problem in Table 8.1.3, your choice *red* is strictly dominated by the randomized choice that assigns probability 0.4 to your choice *blue* and probability 0.6 to your choice *green.* Please verify this. We thus conclude, on the basis of Theorem 2.6.1, that your choice *red* can never be optimal for any belief. In particular, you cannot rationally choose *red* under common belief in rationality.

We will now show, by means of a beliefs diagram, that you can rationally choose the other two colors, *blue* and *green,* under common belief in rationality. Consider the beliefs diagram in Figure 8.1.3. Consider your belief hierarchy that starts at your choice *blue,* in which you believe that, with probability 1, Barbara wears *blue,* and believe, with probability 1, that Barbara believes, with probability 1, that you wear *green.* Clearly, these beliefs yield the second-order expectation that assigns probability 1 to the state $(b, g)$. From Table 8.1.3 it can be seen that your choice *blue* is optimal for this second-order expectation. Or, in other words, your choice *blue* is optimal for the belief hierarchy that starts at your choice *blue.*

In a similar fashion, it can be verified that your choice *green* is optimal for your belief hierarchy that starts at your choice *green,* that Barbara's choice *blue* is optimal for her belief hierarchy that starts at her choice *blue,* and that Barbara's choice *red* is optimal for her belief hierarchy that starts at her choice *red.* Please check this.

Overall, we thus see that in the beliefs diagram, each of the listed choices is optimal for the belief hierarchy that starts at that choice. But then, each of the belief hierarchies in the beliefs diagram expresses common belief in rationality. To see this, consider, for instance, your belief hierarchy that starts at your choice *blue.* In that belief hierarchy, you believe that "Barbara chooses *blue,* Barbara believes that you choose *green,* and Barbara believes that you believe that Barbara chooses *red*". If Barbara indeed believes that you choose *green,* and believes that you believe that Barbara chooses *red,* then it will be optimal for Barbara to choose *blue.* As such, you believe that Barbara chooses rationally.

In the belief hierarchy that starts at your choice *blue,* you also believe that Barbara believes that "you choose *green,* you believe that Barbara chooses *red,* and you believe that Barbara believes that you choose *blue*". If you indeed believe that Barbara chooses *red* and believe that Barbara believes that you choose *blue,* then it will be optimal for you to choose *green.* Hence, you believe that Barbara

believes that you choose rationally. In a similar way, we can verify that the belief hierarchy that starts at your choice *blue* also expresses three-fold belief in rationality, four-fold belief in rationality, and so on. That is, this belief hierarchy expresses common belief in rationality. In the same fashion, it can be checked that also the other three belief hierarchies in the beliefs diagram express common belief in rationality.

As your choice *blue* is optimal for the belief hierarchy that starts at your choice *blue,* your choice *green* is optimal for the belief hierarchy that starts at your choice *green,* and both belief hierarchies express common belief in rationality, we conclude that you can rationally wear *blue* and *green* under common belief in rationality.

It should not be too surprising that under common belief in rationality, you can rationally wear at least two different colors. For suppose that under common belief in rationality you could only rationally wear a single color, say $c$. Then, under common belief in rationality, you must necessarily believe that Barbara believes that you will indeed wear $c$. As such, it would not be possible to surprise Barbara at all by wearing the color $c$, which would make the color $c$ a rather unattractive color to wear in the first place.

## 8.2 Psychological Games

In the previous section we have seen an example of a psychological game, where the preferences over your own choices depended not only on your belief about Barbara's choice, but also on the belief you hold about Barbara's belief about your own choice. That is, the preferences over your own choices depended on your first- and second-order belief. In this section we will provide a general definition of a psychological game, using the example from the previous section as an illustration.

In the example "Surprising Barbara" we have seen that every belief hierarchy of yours induces a *second-order* expectation, which is a probability distribution over pairs $(c_2, c_1)$, where $c_2$ is a choice for Barbara and $c_1$ is a choice for you. In general, a second-order expectation can be defined as follows.

**Definition 8.2.1 (Second-order expectation)** *Consider two players, $i$ and $j$, with sets of choices $C_i$ and $C_j$. A* ***second-order expectation*** *for player $i$ is a probability distribution $e_i$ that assigns to every choice-pair $(c_j, c_i)$, where $c_j \in C_j$ and $c_i \in C_i$, a probability $e_i(c_j, c_i)$.*

Intuitively, the probability $e_i(c_j, c_i)$ indicates the likelihood that player $i$ assigns to the event that "player $j$ chooses $c_j$ and player $j$ believes that player $i$ chooses $c_i$".

When investigating the example "Surprising Barbara", we have seen how we can derive a second-order expectation from the first- and second-order belief of a belief hierarchy, by looking at a (full or partial) beliefs diagram. Moreover, we showed how two different belief hierarchies could generate the same second-order expectation. All this is true in general: From every belief hierarchy we can derive a second-order expectation, and two different belief hierarchies may induce exactly the same second-order expectation.

In the example "Surprising Barbara" we assumed that the preferences over your own choices only depended on the second-order expectation of your belief hierarchy, not on the full specification of your first- and second-order belief. This will be true in general for psychological games as we use them in this book. That is, in a psychological game, player $i$'s preferences over his own choices depends on the second-order expectation, which is a probability distribution over the choice-pairs $(c_j, c_i)$ where $c_j \in C_j$ and $c_i \in C_i$.

This is the same as saying that player $i$'s decision problem is given by (i) the set of choices $C_i$, (ii) the set of states $S_i = C_j \times C_i$, consisting of choice-pairs $(c_j, c_i)$ where $c_j \in C_j$ and $c_i \in C_i$, and (iii) a conditional preference relation $\succsim_i$ that assigns to every belief (second-order expectation) $e_i \in \Delta(S_i)$ some preference relation $\succsim_{i,e_i}$ over his own choices. If we assume, in addition, that the conditional preference relation has an expected utility representation $u_i$, then we arrive at the following general definition of a psychological game with two players.

**Definition 8.2.2 (Psychological game)** *A **psychological game** with two players specifies, for both players $i$, a decision problem $(C_i, S_i, u_i)$, where*

*(i) the set of choices is $C_i$,*

*(ii) the set of states $S_i = C_j \times C_i$ consists of all choice-pairs $(c_j, c_i)$ where $c_j \in C_j$ and $c_i \in C_i$, and*

*(iii) player $i$'s conditional preference relation has an expected utility representation $u_i$, assigning to every choice $c_i \in C_i$ and every state $(c_j, c_i') \in S_i$ some utility $u_i(c_i, (c_j, c_i'))$.*

As an illustration, consider the example "Surprising Barbara" with the decision problems as represented in Table 8.1.3. If we identify you with player 1 and Barbara with player 2, then your decision problem $(C_1, S_1, u_1)$ consists of the following ingredients: (i) Your set of choices is $C_1 = \{$*blue, green, red*$\}$, (ii) your set of states is

$$C_2 \times C_1 = \{(b,b), (b,g), (b,r), (g,b), (g,g), (g,r), (r,b), (r,g), (r,r)\},$$

and your utility function $u_1$ assigns to every choice $c_1$ of yours, and every state $(c_2, c_1')$, where $c_2 \in C_2$ and $c_1' \in C_1$, some utility $u_1(c_1, (c_2, c_1'))$. For instance, $u_1($*blue, (green, blue)*$) = 3$ and $u_1($*green, (green, red)*$) = 2$. Similarly for Barbara.

For the remainder of this and the following chapter, we will assume that there are only two players in a psychological game. As such, the definition of a psychological game above, which restricts to the case of two players, will be sufficient for our purposes.

## 8.3 Common Belief in Rationality

In this section we provide a formal definition of *common belief in rationality* for psychological games. The idea is similar to the other chapters: You believe that your opponent chooses optimally given his second-order expectation, you believe that your opponent believes that you choose optimally given your second-order expectation, and so on. To formally express that you believe that your opponent chooses optimally given his second-order expectation, we need (i) your belief about the opponent's choice (your first-order belief), and (ii) your belief about the opponent's second-order expectation, for which we need your belief about the opponent's belief about your choice (your second-order belief) and your belief about the opponent's belief about your belief about the opponent's choice (your third-order belief). Similarly, to formally express that you believe that your opponent believes that you choose optimally given your second-order expectation, we would need your second-order, third-order and fourth-order belief, and so on. Hence, to formally express common belief in rationality we need your full belief hierarchy.

Contrary to Chapter 5 about *incomplete information* and Chapter 7 about *unawareness,* the notion of a belief hierarchy for psychological games is exactly the same as for *standard* games in Chapter 3.

Indeed, in a psychological game, a belief hierarchy specifies your first-order belief about the opponent's choice, your second-order belief about what the opponent believes about your choice, and so on, exactly like a belief hierarchy for standard games. Consequently, such belief hierarchies can be visualized by means of a beliefs diagram and encoded by means of an epistemic model with types, exactly like we did in Chapter 3 for standard games.

With respect to the epistemic models we use in this part, there is one small difference with Chapter 3: In this part we allow for *infinitely* many types, whereas in Chapter 3 (and in all of the other chapters we have discussed so far) we always assumed a *finite* number of types. This leads to the following, slightly more general, definition of an epistemic model.

**Definition 8.3.1 (Epistemic model)** *An* ***epistemic model*** $M = (T_i, b_i)_{i \in I}$ *for a psychological game specifies*

*(a) for every player $i$ a set of types $T_i$, which may contain infinitely many types, and*

*(b) for every player $i$ and every type $t_i \in T_i$, a probability distribution $b_i(t_i)$ on the opponents' choice-type combinations. We assume that $b_i(t_i)$ assigns positive probability to only finitely many opponents' choice-type combinations. This probability distribution $b_i(t_i)$ represents $t_i$'s belief about the opponents' choices and types.*

The reason we allow for infinitely many types will become clear in Section 8.4 when we discuss the recursive procedure.

We will see that the definition of common belief in rationality is almost exactly the same as in Chapter 3 for standard games. The only difference lies in the way we define an optimal choice for a type in an epistemic model: For standard games, a choice was called optimal for a type if it is optimal for the *first-order belief* this type has. In a psychological game, a choice is called optimal for a type if it is optimal for the *second-order expectation* this type has. Apart from this, the definition of common belief in rationality will be identical to the one from Chapter 3.

### 8.3.1 Optimal Choices for Types

Consider an epistemic model, and a type $t_i$ for player $i$ within this epistemic model. In Chapter 3 we have seen how we can derive a full belief hierarchy for such a type. In particular, type $t_i$ will then induce a second-order expectation. We say that a choice is optimal for the type $t_i$ if it is optimal given $t_i$'s second-order expectation.

**Definition 8.3.2 (Optimal choice for type)** *Consider a type $t_i$ for player $i$ in an epistemic model, and suppose that $t_i$ has the second-order expectation $e_i$. Then, the choice $c_i$ is* ***optimal*** *for the type $t_i$ if*

$$\sum_{(c_j, c_i') \in S_i} e_i(c_j, c_i') \cdot u_i(c_i, (c_j, c_i')) \geq \sum_{(c_j, c_i') \in S_i} e_i(c_j, c_i') \cdot u_i(c_i'', (c_j, c_i'))$$

*for all $c_i'' \in C_i$.*

That is, choice $c_i$ yields the highest possible expected utility, given the second-order expectation that type $t_i$ has. In the sequel, we will write

$$u_i(c_i, t_i) := \sum_{(c_j, c_i') \in S_i} e_i(c_j, c_i') \cdot u_i(c_i, (c_j, c_i'))$$

| | | | |
|---|---|---|---|
| **Types** | $T_1 = \{t_1^{blue}, t_1^{green}\}, \quad T_2 = \{t_2^{blue}, t_2^{red}\}$ | | |
| **Beliefs for you** | $b_1(t_1^{blue})$ | $=$ | $(0.8) \cdot (blue,\, t_2^{blue}) + (0.2) \cdot (red,\, t_2^{red})$ |
| | $b_1(t_1^{green})$ | $=$ | $(red,\, t_2^{red})$ |
| **Beliefs for Barbara** | $b_2(t_2^{blue})$ | $=$ | $(green,\, t_1^{green})$ |
| | $b_2(t_2^{red})$ | $=$ | $(0.9) \cdot (blue,\, t_1^{blue}) + (0.1) \cdot (green,\, t_1^{green})$ |

Table 8.3.1 Epistemic model for "Surprising Barbara"

to denote the expected utility that player $i$ will have when he chooses $c_i$ and holds the belief hierarchy encoded by $t_i$.

To illustrate the notion of an optimal choice for a type, consider the example "Surprising Barbara" and the associated epistemic model in Table 8.3.1. Let us concentrate on your type $t_1^{blue}$. We will show that your choice *blue* is optimal for the type $t_1^{blue}$.

To verify this, we first need to derive the second-order expectation that type $t_1^{blue}$ has. Note that type $t_1^{blue}$ assigns probability 0.8 to the event that "Barbara chooses *blue* and has type $t_2^{blue}$ " and probability 0.2 to the event that "Barbara chooses *red* and has type $t_2^{red}$ ". In turn, Barbara's type $t_2^{blue}$ assigns probability 1 to you choosing *green,* whereas Barbara's type $t_2^{red}$ assigns probability 0.9 to you choosing *blue* and probability 0.1 to you choosing *green.* By putting this together, we see that type $t_1^{blue}$ assigns probability 0.8 to the event that "Barbara chooses *blue* and assigns probability 1 to you choosing *green* ", and probability 0.2 to the event that "Barbara chooses *red* and assigns probabilities 0.9 and 0.1 to you choosing *blue* and *green,* respectively".

As such, the second-order expectation $e_1$ induced by type $t_1^{blue}$ is

$$\begin{aligned} e_1 &= (0.8) \cdot 1 \cdot (b, g) + (0.2) \cdot (0.9) \cdot (r, b) + (0.2) \cdot (0.1) \cdot (r, g) \\ &= (0.8) \cdot (b, g) + (0.18) \cdot (r, b) + (0.02) \cdot (r, g). \end{aligned}$$

The expected utilities that you obtain by making your three choices are thus

$$\begin{aligned} u_1(blue,\, t_1) &= (0.8) \cdot 3 + (0.18) \cdot 3 + (0.02) \cdot 6 = 3.06, \\ u_1(green,\, t_1) &= (0.8) \cdot 2 + (0.18) \cdot 4 + (0.02) \cdot 2 = 2.36 \text{ and} \\ u_1(red,\, t_1) &= (0.8) \cdot 2 + (0.18) \cdot 1 + (0.02) \cdot 1 = 1.8. \end{aligned}$$

As *blue* yields the highest expected utility, your choice *blue* is optimal for the type $t_1^{blue}$.

**Question 8.3.1** *Show, in a similar way, that the choice green is optimal for your type* $t_1^{green}$.

It may also be verified that the choice *blue* is optimal for Barbara's type $t_2^{blue}$, and that the choice *red* is optimal for Barbara's type $t_2^{red}$. Please check this.

### 8.3.2 Definition of Common Belief in Rationality

Recall how we defined common belief in rationality in Chapter 3, for standard games. We started by defining what it means for a type $t_i$ to believe in the opponent's rationality. Formally, it meant that $t_i$

only assigns positive probability to opponents' choice-type pairs $(c_j, t_j)$ where the choice $c_j$ is optimal for the type $t_j$. This definition can literally be translated to the context of psychological games.

**Definition 8.3.3 (Belief in the opponent's rationality)** *Type $t_i$ **believes in the opponent's rationality** if $b_i(t_i)$ only assigns positive probability to choice-type pairs $(c_j, t_j)$ where the choice $c_j$ is optimal for the type $t_j$.*

However, as we have seen above, optimal choices for a type are defined differently than for standard games.

In Chapter 3, we then recursively defined $k$-fold belief in rationality for all $k \in \{1, 2, 3, ...\}$, which finally enabled us to define common belief in rationality. Also this definition can be translated without any change to the class of psychological games.

**Definition 8.3.4 (Common belief in rationality)** *A type $t_i$ expresses 1-fold belief in rationality if $t_i$ believes in the opponent's rationality.*

*A type $t_i$ expresses 2-fold belief in rationality if $b_i(t_i)$ only assigns positive probability to opponent's types that express 1-fold belief in rationality.*

*A type expresses 3-fold belief in rationality if $b_i(t_i)$ only assigns positive probability to opponent's types that express 2-fold belief in rationality.*

*And so on.*

*A type $t_i$ expresses **common belief in rationality** if it expresses 1-fold belief in rationality, 2-fold belief in rationality, 3-fold belief in rationality, and so on, ad infinitum.*

Finally, like in Chapter 3, we say that you can rationally make a choice $c_i$ under common belief in rationality if there is a belief hierarchy that expresses common belief in rationality and supports the choice $c_i$.

**Definition 8.3.5 (Rational choice under common belief in rationality)** *Player $i$ can **rationally make choice** $c_i$ **under common belief in rationality** if there is some epistemic model $M = (T_i, b_i)_{i \in I}$, and some type $t_i \in T_i$ for player $i$ within that model, such that (a) type $t_i$ expresses common belief in rationality, and (b) choice $c_i$ is optimal for the type $t_i$.*

As an illustration, consider the example "Surprising Barbara", and the associated epistemic model in Table 8.3.1. We have seen that your choice *blue,* your choice *green,* Barbara's choice *blue* and Barbara's choice *red* are optimal for the types $t_1^{blue}$, $t_1^{green}$, $t_2^{blue}$ and $t_2^{red}$, respectively. But then, it may be verified that each of the types in the epistemic model believes in the opponent's rationality. Can you explain why?

Since every type in the model believes in the opponent's rationality, we can conclude, in a similar way as in earlier chapters, that every type in the model also expresses *common* belief in rationality. As your choice *blue* is optimal for your type $t_1^{blue}$ and your choice *green* is optimal for your type $t_1^{green}$, it follows that you can rationally wear *blue* and *green* under common belief in rationality. In fact, we have already seen this in Section 8.1, where we argued in terms of beliefs diagrams instead of types.

| **You** | $(b,b)$ | $(b,g)$ | $(b,r)$ | $(g,b)$ | $(g,g)$ | $(g,r)$ | $(r,b)$ | $(r,g)$ | $(r,r)$ |
|---|---|---|---|---|---|---|---|---|---|
| *blue* | 0 | 3 | 3 | 3 | 6 | 6 | 3 | 6 | 6 |
| *green* | 4 | 2 | 4 | 2 | 0 | 2 | 4 | 2 | 4 |

| **Barbara** | $(b,b)$ | $(b,g)$ | $(b,r)$ | $(g,b)$ | $(g,g)$ | $(g,r)$ | $(r,b)$ | $(r,g)$ | $(r,r)$ |
|---|---|---|---|---|---|---|---|---|---|
| *blue* | 0 | 2 | 2 | 2 | 4 | 4 | 2 | 4 | 4 |
| *red* | 6 | 6 | 3 | 6 | 6 | 3 | 3 | 3 | 0 |

Table 8.4.1 One-fold reduced decision problems for "Surprising Barbara"

## 8.4 Recursive Procedure

As a next step we will develop a recursive elimination procedure that yields exactly those choices that the players can rationally make under common belief in rationality. We start by a procedure that recursively eliminates choices and states at the players' decision problems, like we did in the earlier chapters. Although this procedure works for the example "Surprising Barbara", we will show by means of another example that the procedure may fail to eliminate certain choices that cannot be rationally made under common belief in rationality. That is, elimination of choices and states may not be enough to capture the choices that are possible under common belief in rationality. As a remedy to this problem we develop a more sophisticated procedure that eliminates second-order expectations, and show that this procedure *does* yield precisely those choices that are possible under common belief in rationality.

### 8.4.1 Elimination of Choices and States

Let us go back to the example "Surprising Barbara" with the decision problems as depicted in Table 8.1.3. Can we capture the consequences of common belief in rationality by means of a recursive elimination procedure that removes choices and states? Recall that your choice *red* is strictly dominated by the randomized choice $(0.4)\cdot$ *blue* + $(0.6)\cdot$ *green,* and that Barbara's choice *green* is strictly dominated by the randomized choice $(0.6)\cdot$ *blue* + $(0.4)\cdot$ *red.* By Theorem 2.6.1 we can thus conclude that your choice *red* and Barbara's choice *green* are never optimal for any belief about the states. In other words, these choices are never optimal for any second-order expectation. As such, we can safely remove these choices from the two decision problems, and obtain the one-fold reduced decision problems in Table 8.4.1.

Consider your one-fold reduced decision problem. If you express 1-fold belief in rationality – that is, believe in Barbara's rationality – then you must believe that Barbara will not choose *green.* Moreover, if you express 2-fold belief in rationality, which means that you believe that Barbara believes in your rationality, then you must believe that Barbara believes that you will not choose *red.* As such, you can discard all states $(c_2, c_1)$ that start with Barbara's choice *green* or end with your choice *red.* We may thus eliminate these states from your decision problem, as you will assign probability 0 to these states if you express up to 2-fold belief in rationality. In the reduced decision problem so obtained, none of your two remaining choices is strictly dominated.

By a similar argument, we can eliminate from Barbara's decision problem all states $(c_1, c_2)$ that start with your choice *red* or end with Barbara's choice *green.* We then obtain a reduced decision problem where no remaining choice for Barbara is strictly dominated. We thus arrive at the two-fold reduced decision problems in Table 8.4.2.

| **You** | $(b,b)$ | $(b,g)$ | $(r,b)$ | $(r,g)$ |
|---|---|---|---|---|
| *blue* | 0 | 3 | 3 | 6 |
| *green* | 4 | 2 | 4 | 2 |

| **Barbara** | $(b,b)$ | $(b,r)$ | $(g,b)$ | $(g,r)$ |
|---|---|---|---|---|
| *blue* | 0 | 2 | 2 | 4 |
| *red* | 6 | 3 | 6 | 3 |

Table 8.4.2 Two-fold reduced decision problems for "Surprising Barbara"

| **You** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 0 | 0 | 0 | 1 |
| *white* | 1 | 1 | 1 | 1 |

| **Barbara** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 1 | 1 | 1 | 1 |
| *white* | 1 | 0 | 0 | 0 |

Table 8.4.3 Decision problems for "The black and white dinner"

Recall, from our discussion in Section 8.1, that under common belief in rationality you can rationally choose *blue* and *green,* whereas Barbara can rationally choose *blue* and *red.* Note that these are precisely the choices that survived the procedure above.

However, as we will see next, we have been "lucky" here: In general, the procedure used above may be too permissive, and fail to eliminate certain choices that cannot rationally be made under common belief in rationality.

### 8.4.2 Why Elimination of Choices and States is Not Enough

As announced above, we will now introduce a new example where the procedure above *fails* to characterize the choices that can rationally be made under common belief in rationality.

**Example 8.2: The black and white dinner.**

You and Barbara have been invited to Chris' house for a nice dinner. Strangely perhaps, Chris imposes one condition: You must either come completely dressed in *black* or completely dressed in *white*, and similarly for Barbara. The question is: Should you go in *black* or in *white*?

You have a clear preference for *white*, but there is one exception: If you believe, with probability 1, that by wearing *black* you wear a different color than Barbara and happen to surprise Barbara by the color you wear, then you would be indifferent between *black* and *white*. In all other cases, you prefer *white* to *black.* Barbara's conditional preference relation is similar, but she has a clear preference for *black* instead of *white.* This story can be captured by the decision problems in Table 8.4.3. Here, the state $(b,w)$ in your decision problem represents the event where Barbara wears *black* and believes that you wear *white.* Similarly for the other states.

We will first show that under common belief in rationality, you can only rationally wear *white.* To see why, note that wearing *black* can only be optimal for you if in your second-order expectation you assign probability 1 to the state $(w,w)$. Indeed, in that case you would be indifferent between *black* and *white,* but as soon as you assign some positive probability to a state different from $(w,w)$ you would prefer *white* to *black.*

In turn, your second-order expectation will only assign probability 1 to the state $(w,w)$ if you believe, with probability 1, that Barbara wears *white,* and believe, with probability 1, that Barbara assigns probability 1 to you wearing *white* as well. However, if Barbara assigns probability 1 to you

wearing *white,* then Barbara should only assign positive probability to the states $(w, b)$ and $(w, w)$. But then, it can be seen from Barbara's decision problem that she would prefer *black* to *white.* Hence, if you believe in Barbara's rationality, then you cannot assign probability 1 to the event that Barbara wears *white* and that Barbara assigns probability 1 to you wearing *white.* But then, as we have seen above, it cannot be optimal for you to wear *black.* In particular, under common belief in rationality you cannot rationally wear *black.*

However, your choice *black* cannot be eliminated by the recursive procedure we outlined above. Indeed, in your and Barbara's decision problem, no choice is strictly dominated, and therefore we cannot eliminate any choice in the first round. As a consequence, no states can be eliminated in the next round either, and the procedure terminates immediately. As such, the procedure fails to eliminate your choice *black,* which cannot rationally be made under common belief in rationality.

This raises the question: What is wrong with this procedure? The problem is that the elimination of states is not sufficiently fine-grained to capture all the consequences of common belief in rationality. To see why, let us return to the example "The black and white dinner" above. Since no choice is strictly dominated in the first round, we would eliminate no states in the second round. Indeed, we would only eliminate a state $(c_2, c_1)$ for you if either the choice $c_2$ for Barbara, or your own choice $c_1$, was strictly dominated in the previous round. Similarly for Barbara.

However, consider the state $(w, w)$ for you, where Barbara chooses *white* while believing that you choose *white* as well. From Barbara's decision problem we see that it can never be optimal for Barbara to wear *white* if she believes, with some positive probability, that you wear *white.* Can you explain this? But then, if you believe in Barbara's rationality, your second-order expectation should not assign a positive probability to the state $(w, w)$, because otherwise you would assign a positive probability to the event that "Barbara chooses *white* while assigning a positive probability to you choosing *white*". This means, in turn, that we should discard the state $(w, w)$ from consideration in your decision problem. But then, your choice *black* would no longer be optimal for any remaining belief, and could thus be eliminated.

Hence, the procedure discussed above does not rule out the state $(w, w)$ for you, whereas it should be discarded on the basis of common belief in rationality. In that sense, the procedure was not sufficiently fine-grained.

### 8.4.3 Elimination of Second-Order Expectations

Let us return to the example "The black and white dinner". We have seen that you cannot rationally wear *black* under common belief in rationality. At the same time, a procedure that eliminates choices and states based on strict dominance alone is not sufficiently fine-grained to eliminate your choice *black.* The problem is that such a procedure is not able to rule out the state $(w, w)$ in your decision problem. Indeed, since Barbara's choice *white* is not strictly dominated in her decision problem, you may believe that Barbara wears *white,* and hence such a procedure would allow for states that start with $w$. On the other hand, your choice *white* is not strictly dominated in your decision problem, and hence you may believe that Barbara believes that you wear *white.* Therefore, such a procedure would allow for states that end with $w$. In particular, the procedure would allow for the state $(w, w)$.

In a sense, the procedure views the two components in the state $(c_2, c_1)$ as disconnected: We would only remove the state $(c_2, c_1)$ if either $c_2$ is strictly dominated in Barbara's decision problem, or $c_1$ is strictly dominated in your decision problem, but we do not pay any attention to the connection between $c_1$ and $c_2$. But the connection between the choices $c_1$ and $c_2$ in the state $(c_2, c_1)$ is crucial if we reason in accordance with common belief in rationality. Indeed, the state $(c_2, c_1)$ describes the

| **You** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 0 | 0 | 0 | 3 |
| *white* | 2 | 2 | 2 | 2 |

| **Barbara** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 2 | 2 | 2 | 2 |
| *white* | 3 | 0 | 0 | 0 |

Table 8.4.4 Decision problems for "The black and white dinner with a twist"

event where Barbara chooses $c_2$ *and* believes that you choose $c_1$. Since you believe Barbara to choose optimally given her belief, her choice $c_2$ must be justified by the belief she holds about you.

Suppose now that the choice $c_2$ for Barbara is optimal for some belief, but not for any belief that assigns a positive probability to you choosing $c_1$. If you believe in Barbara's rationality, then your second-order expectation should assign probability zero to the state $(c_2, c_1)$. For if you were to assign a positive probability to the state $(c_2, c_1)$, then you would assign a positive probability to the event that "Barbara chooses $c_2$ and assigns a positive probability to you choosing $c_1$". This, however, would violate your belief in Barbara's rationality.

In "The black and white dinner", for instance, the choice *white* for Barbara is optimal for some belief, but not for any belief that assigns a positive probability to you choosing *white.* For that reason, your second-order expectation must assign probability zero to the state $(w, w)$ in your decision problem. But then, the only optimal choice for you is to wear *white.*

Hence, a procedure that captures the implications of common belief in rationality should rule out the state $(w, w)$ in your decision problem. Not because Barbara's choice *white* is never optimal for any belief, nor because your choice *white* is never optimal for any belief, but because Barbara cannot rationally choose *white* if she assigns a positive probability to you choosing *white.* We should thus take seriously the connection between Barbara's choice *white* and Barbara's belief that you choose *white* in the state $(w, w)$.

Ruling out the state $(w, w)$ in your decision problem naturally restricts the second-order expectations that you may hold: From that moment on, we would only consider second-order expectations of yours that assign probability zero to the ruled out state $(w, w)$. However, there are other psychological games where common belief in rationality leads us to consider *restricted* sets of second-order expectations, but where these restrictions *cannot* be translated into the *elimination* of states. The following example will illustrate this.

**Example 8.3: The black and white dinner with a twist.**

It is a few weeks later, and Chris has again invited you and Barbara for a *black and white dinner* at his house. You still have a strong preference for *white* compared to *black*, but if you believe, with probability 1, that Barbara wears *white,* and believe, with probability 1, that Barbara believes, with probability 1, that you wear *white,* then you slightly prefer wearing *black* to wearing *white.* Similarly for Barbara, who has a strong preference for *black* compared to *white.* These new conditional preference relations can be captured by the decision problems in Table 8.4.4.

Which color(s) can you rationally wear under common belief in rationality? To answer this question in a systematic way, we first graphically depict your conditional preference relation in the left-hand panel of Figure 8.4.1. Note that we have four states, and hence we can identify every second-order expectation (which is a probability distribution over these four states) with the points in a pyramid, like we did in Figure 2.4.4. The extreme points of the pyramid correspond to the four states. The vector $(1/3, 0, 0, 2/3)$, for instance, corresponds to the second-order expectation that assigns probabilities $1/3, 0, 0$ and $2/3$ to the states $(w, b), (b, w), (b, b)$ and $(w, w)$, respectively. Hence, we number the states in the order $(w, b), (b, w), (b, b)$ and $(w, w)$.

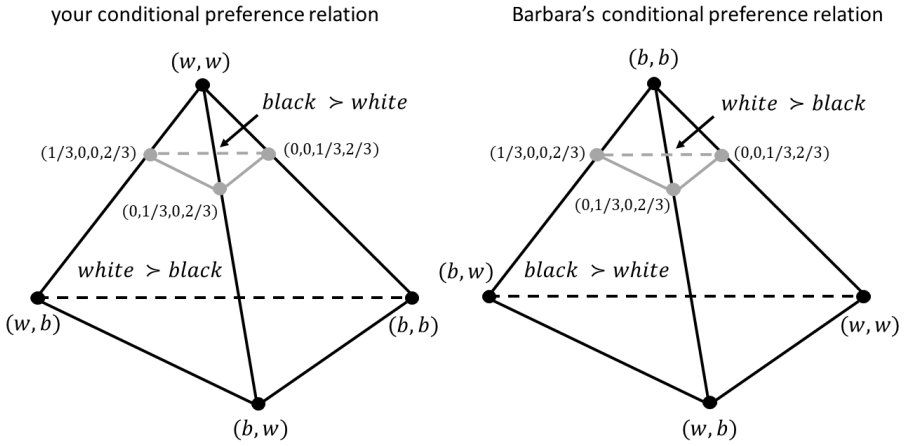



Figure 8.4.1 Graphical representation of conditional preference relations in "The black and white dinner with a twist"

From your decision problem in Table 8.4.4 we see that your expected utility from wearing *black* is

$$u_1(black) = 3 \cdot e_1(w, w),$$

where $e_1(w, w)$ is the probability that your second-order expectation $e_1$ assigns to the state $(w, w)$. Moreover, the expected utility from wearing *white* is simply

$$u_1(white) = 2.$$

Hence, you prefer *black* to *white* precisely when $e_1(w, w) > 2/3$, that is, when the probability that your second-order expectation assigns to the state $(w, w)$ is more than $2/3$. These are exactly the second-order expectations above the grey triangle in the left-hand panel of Figure 8.4.1.

As such, you prefer *white* to *black* for all second-order expectations below the grey triangle, and you are indifferent between the two choices for all second-order expectations on the grey triangle.

In a similar way, we can graphically depict Barbara's conditional preference relation in the right-hand panel of Figure 8.4.1. For this representation, we have ordered her states by $(b, w), (w, b), (w, w)$ and $(b, b)$, so that the vector $(1/3, 0, 0, 2/3)$ corresponds to the second-order expectation that assigns probabilities $1/3, 0, 0$ and $2/3$ to the states $(b, w), (w, b), (w, w)$ and $(b, b)$, respectively.

From Barbara's conditional preference relation we can conclude that she can only rationally choose *white* if her second-order expectation assigns probability at least $2/3$ to the state $(b, b)$. In particular, this means that she can only rationally choose *white* if she assigns probability at least $2/3$ to you choosing *black.* We say that her choice *white* is supported by the set of first-order beliefs that assign probability at least $2/3$ to you choosing *black.* Or, written down more formally,

$$B_2(white) = \{b_2 \in \Delta(C_1) \mid b_2(black) \geq 2/3\}, \tag{8.4.1}$$

where $B_2(white)$ denotes the set of first-order beliefs for Barbara that supports her choice *white,* and $\Delta(C_1)$ denotes the set of all first-order beliefs. Here, the subindex 2 refers to player 2 (Barbara).

On the other hand, Barbara's choice *black* is supported by *every* first-order belief about your choice. Indeed, it can be seen from the right-hand panel of Figure 8.4.1 that Barbara's choice *black*

is optimal for every second-order expectation that assigns positive probability to the states $(b, w)$ and $(w, b)$ only. Since every first-order belief for Barbara is part of such a second-order expectation, we conclude that Barbara's choice *black* is supported by every first-order belief about your choice. Or, more formally,

$$B_2(black) = \Delta(C_1), \tag{8.4.2}$$

where $B_2(black)$ denotes the set of first-order beliefs that support Barbara's choice *black,* and $\Delta(C_1)$ denotes the set of all first-order beliefs.

Now suppose that you believe in Barbara's rationality. Then, whenever you assign a positive probability to Barbara choosing *white,* you must believe that Barbara assigns probability at least 2/3 to you choosing *black.* What does this mean for your second-order expectation $e_1$?

Let us denote by

$$e_1(w_1 \mid w_2) := \frac{e_1(w_2, w_1)}{e_1(w_2, w_1) + e_1(w_2, b_1)} \tag{8.4.3}$$

the probability by which you believe that Barbara believes that you choose *white,* conditional on the event that Barbara chooses *white.* Here, we write the subindex 1 at the choices $w_1$ and $b_1$ to indicate that this is a choice for you (player 1), whereas we write the subindex 2 at the choice $w_2$ to stress that this choice is for Barbara (player 2). Note that $e_1(w_2, w_1) + e_1(w_2, b_1)$ is the total probability you assign to Barbara wearing *white,* and hence the fraction above indeed represents the probability by which you believe that Barbara believes that you choose *white,* conditional on the event that Barbara chooses *white.* Similarly,

$$e_1(b_1 \mid w_2) := \frac{e_1(w_2, b_1)}{e_1(w_2, w_1) + e_1(w_2, b_1)} \tag{8.4.4}$$

is the probability by which you believe that Barbara believes that you choose *black,* conditional on Barbara choosing *white.* Note that the two probabilities $e_1(w_1 \mid w_2)$ and $e_1(b_1 \mid w_2)$ together constitute a first-order belief for Barbara about your choices *white* and *black.* We denote this first-order belief by $e_1(\cdot \mid w_2)$. It represents the first-order belief you believe Barbara to have, conditional on Barbara wearing *white.*

Recall from above that, whenever you assign a positive probability to Barbara choosing *white,* you must believe that Barbara assigns probability at least 2/3 to you choosing *black.* In view of (8.4.4), this means that

$$e_1(b_1 \mid w_2) \geq 2/3$$

whenever you assign a positive probability to Barbara wearing *white.* Together with (8.4.1) we then conclude that

$$e_1(\cdot \mid w_2) \text{ must be in } B_2(white) \tag{8.4.5}$$

whenever you assign a positive probability to Barbara choosing *white.* That is, if you assign a positive probability to Barbara wearing *white,* then the first-order belief you believe Barbara to have conditional on her wearing *white* must support her choice *white.*

By a similar argument, we conclude that

$$e_1(\cdot \mid b_2) \text{ must be in } B_2(black) \tag{8.4.6}$$

whenever you assign a positive probability to Barbara wearing *black.* However, since, by (8.4.2), $B_2(black)$ contains all first-order beliefs for Barbara, this imposes no extra conditions on your second-order expectations.

Summarizing, we see that if you believe in Barbara's rationality, then your second-order expectation $e_1$ must satisfy the requirements (8.4.5) and (8.4.6). The second-order expectations that meet these

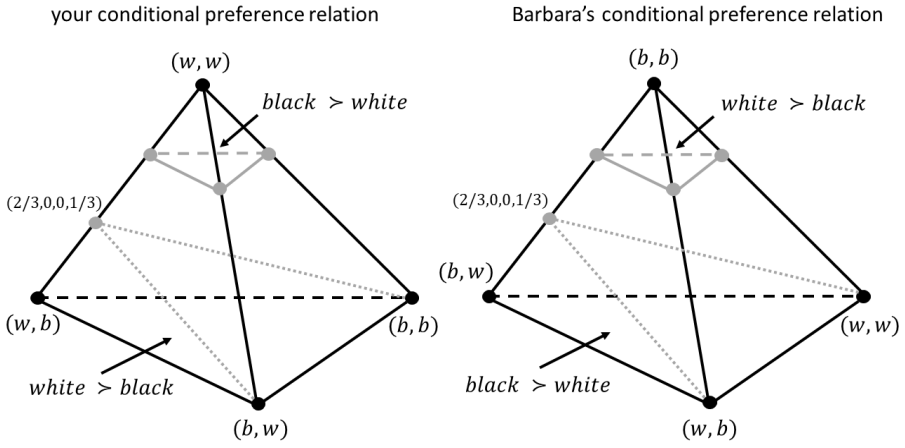



Figure 8.4.2 Second-order expectations consistent with one-fold belief in rationality

conditions are depicted in the left-hand panel of Figure 8.4.2, below the dotted grey triangle. Hence, these are exactly the second-order expectations for you that are consistent with one-fold belief in rationality.

Note that the set of second-order expectations below the dotted grey triangle has as the extreme points the second-order expectation $(2/3, 0, 0, 1/3)$ and the states $(w, b), (b, w)$ and $(b, b)$.

**Question 8.4.1** *Show that each of these four extreme second-order expectations satisfies the conditions (8.4.5) and (8.4.6).*

Note that the states $(w, b), (b, w)$ and $(b, b)$ correspond to the second-order expectations $(1, 0, 0, 0)$, $(0, 1, 0, 0)$ and $(0, 0, 1, 0)$ respectively. Now, take an arbitrary second-order expectation $e_1$ below the dotted grey triangle. Then, $e_1$ can be written as

$$e_1 = \lambda_1 \cdot (2/3, 0, 0, 1/3) + \lambda_2 \cdot (1, 0, 0, 0) + \lambda_3 \cdot (0, 1, 0, 0) + \lambda_4 \cdot (0, 0, 1, 0),$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are some numbers that are greater than, or equal to, zero, and where $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$.

**Question 8.4.2** *Show that every such second-order expectation $e_1$ satisfies the conditions (8.4.5) and (8.4.6).*

We thus conclude that all second-order expectations below the dotted grey triangle satisfy the conditions (8.4.5) and (8.4.6), and are thus consistent with one-fold belief in rationality. Hence, the second-order expectations that are consistent with one-fold belief in rationality are precisely those that lie below the dotted grey triangle. Note that these second-order expectations are not obtained by eliminating a state, which is what happened in the example "The black and white dinner". Indeed, every state in your decision problem may receive positive probability by a second-order expectation that lies below the lower grey triangle, and thus no state can be completely discarded if you believe in Barbara's rationality.

Now focus on your second-order expectations that are consistent with one-fold belief in rationality. As all of these second-order expectations lie below the upper grey triangle, it follows from your


Figure 8.4.3 Beliefs diagram for "The black and white dinner with a twist"

conditional preference relation that under one-field belief in rationality you will always prefer *white* to *black.* That is, if you believe in Barbara's rationality, then your only optimal choice is wearing *white.*

We can can do a similar analysis for Barbara, and conclude that Barbara's second-order expectations that are consistent with one-fold belief in rationality are exactly those that lie below the dotted grey triangle in the right-hand panel of Figure 8.4.2. Please verify this. As all of these second-order expectations lie below the upper grey triangle, we conclude from Barbara's conditional preference relation that she will always prefer *black* to *white* if she believes in your rationality. In particular, under one-fold belief in rationally, Barbara can only rationally wear *black.*

We thus see that under common belief in rationality, the only colors that can possibly be chosen rationally are *white* for you and *black* for Barbara. The beliefs diagram in Figure 8.4.3 shows that you can indeed rationally wear *white,* and Barbara can indeed rationally wear *black,* under common belief in rationality. Your belief hierarchy in that beliefs diagram states that you believe, with probability 1, that Barbara wears *black,* that you believe, with probability 1, that Barbara believes, with probability 1, that you wear *white,* that you believe, with probability 1, that Barbara believes, with probability 1, that you believe, with probability 1, that Barbara wears *black,* and so on.

It is easily seen that this belief hierarchy believes in Barbara's rationality. Indeed, you believe in the event that "Barbara chooses *black,* Barbara believes that you choose *white,* and Barbara believes that you believe that Barbara chooses *black*". From Table 8.4.4 we know that wearing *black* is optimal for Barbara if she believes that you wear *white* and believes that you believe that Barbara wears *black.* As such, you believe in Barbara's rationality. By similar arguments, it can be verified that your belief hierarchy also expresses 2-fold belief in rationality, 3-fold belief in rationality, and so on, and hence we conclude that your belief hierarchy expresses common belief in rationality. As you believe that Barbara chooses *black,* and believe that Barbara believes that you choose *white,* your choice *white* is optimal for you under this belief hierarchy. Hence, you can rationally wear *white* under common belief in rationality.

In a similar fashion, it can be shown that Barbara's belief hierarchy in this beliefs diagram also expresses common belief in rationality. As wearing *black* is optimal for Barbara under this belief hierarchy, we know that Barbara can rationally wear *black* under common belief in rationality.

### 8.4.4 One-Fold Belief in Rationality

Based on our insights from the example "The black and white dinner with a twist" we will now show how to characterize, in general, those second-order expectations that are consistent with one-fold belief in rationality, two-fold belief in rationality, and so on. And we will use this to develop a general elimination procedure that characterizes those choices that can rationally be made under common belief in rationality.

As usual, we start by characterizing those choices that are optimal for *some* belief, without imposing any restrictions yet on the belief. Consider the decision problem for player $i$, where the states

are the choice-pairs $(c_j, c_i)$ consisting of a choice $c_j$ for player $j$ and a choice $c_i$ for player $i$. Moreover, by the definition of a psychological game, player $i$'s preferences depend exclusively on $i$'s second-order expectation, that is, his belief about the states in the decision problem. But then, we know by Theorem 2.6.1 that a choice for player $i$ is optimal for some second-order expectation precisely when this choice is not strictly dominated in his decision problem. We thus may start, in round 1, by eliminating those choices for the two players that are strictly dominated in their respective decision problems.

As a next step we characterize those choices that player $i$ can rationally make if he expresses 1-fold belief in rationality. Suppose that player $i$ believes in opponent $j$'s rationality. Like we did in the example "The black and white dinner with a twist", we can identify for every choice $c_j$ of player $j$ that has not been eliminated in round 1 the set $B_j^2(c_j)$ of first-order beliefs that support the choice $c_j$. Formally, $B_j^2(c_j)$ contains those first-order beliefs $b_j$ for player $j$ such that there is a second-order expectation $e_j$ for which the choice $c_j$ is optimal and that induces the first-order belief $b_j$.

If player $i$ believes in $j$'s rationality then, whenever he assigns a positive probability to an opponent's choice $c_j$, he must believe that $j$ holds a first-order belief in $B_j^2(c_j)$ that supports this choice. As we have seen in the example "The black and white dinner with a twist", this means that $i$'s second-order expectation $e_i$ must satisfy the following condition for every choice $c_j$ that $e_i$ assigns positive probability to:

$$e_i(\cdot \mid c_j) \text{ must be in } B_j^2(c_j), \tag{8.4.7}$$

where $e_i(\cdot \mid c_j)$ is the first-order belief that $i$ believes $j$ to have, conditional on $j$ choosing $c_j$. Formally, $e_i(\cdot \mid c_j)$ is the first-order belief for player $j$ where

$$e_i(c_i \mid c_j) := \frac{e_i(c_j, c_i)}{\sum_{c_i' \in C_i} e_i(c_j, c_i')} \text{ for every choice } c_i \in C_i.$$

Let us denote by $E_i^2$ the set of second-order expectations $e_i$ for player $i$ that satisfy condition (8.4.7). Then, by our argument above, $E_i^2$ contains precisely those second-order expectations for player $i$ that are consistent with one-fold belief in rationality.

Hence, the choices that player $i$ can rationally make under 1-fold belief in rationality are precisely those choices that are optimal for some second-order expectation in $E_i^2$. We may thus eliminate all choices for player $i$ that are not optimal for any second-order expectation in $E_i^2$.

Summarizing, round 2 of the procedure would thus look as follows: In $i$'s decision problem, we identify for every choice $c_j$ that was not eliminated for player $j$ in round 1 the set $B_j^2(c_j)$ of first-order beliefs that support the choice $c_j$. On the basis of these sets, we then find the set $E_i^2$ of second-order expectations for player $i$ that satisfy condition (8.4.7). We finally eliminate for player $i$ those choices that are not optimal for any second-order expectation in $E_i^2$. This concludes round 2. The choices for player $i$ that survive round 2 are precisely the choices that player $i$ can rationally make under 1-fold belief in rationality.

### 8.4.5 Two-Fold Belief in Rationality

We will now characterize the choices that player $i$ can rationally make if he expresses 1-fold and 2-fold belief in rationality. Suppose that player $i$ expresses 1-fold and 2-fold belief in rationality. Then, player $i$ believes that player $j$ chooses rationally and that player $j$ expresses 1-fold belief in rationality. We have seen above that player $j$ must hold a second-order expectation in $E_j^2$ if he expresses 1-fold belief in rationality. As such, player $i$ should only assign positive probability to an opponent's choice $c_j$ if $c_j$ is optimal for some second-order expectation in $E_j^2$.

For every opponent's choice $c_j$, let $B_j^3(c_j)$ be the set of first-order beliefs $b_j \in \Delta(C_i)$ for player $j$ such that there is a second-order expectation $e_j$ in $E_j^2$ where (i) $e_j$ has the first-order belief $b_j$, and (ii) the choice $c_j$ is optimal for the second-order expectation $e_j$. Hence, $B_j^3(c_j)$ are the first-order beliefs for player $j$ that support the choice $c_j$ and that are consistent with 1-fold belief in rationality.

Then, by our argument above we conclude that, whenever player $i$ assigns a positive probability to an opponent's choice $c_j$, he must believe that player $j$ holds a first-order belief in $B_j^3(c_j)$ to support that choice. Similarly to what we have seen above, this means that player $i$ must hold a second-order expectation $e_i$ where

$$e_i(\cdot \mid c_j) \text{ must be in } B_j^3(c_j) \tag{8.4.8}$$

for every choice $c_j$ that receives a positive probability by $e_i$. Let us denote by $E_i^3$ the set of second-order expectations for player $i$ that satisfy condition (8.4.8). Then, $E_i^3$ contains precisely those second-order expectations that are consistent with 1-fold and 2-fold belief in rationality.

If player $i$ chooses rationally given his second-order expectation, then he must make a choice that is optimal for some second-order expectation in $E_i^3$. We may thus eliminate all choices for player $i$ that are not optimal for any second-order expectation in $E_i^3$. This would conclude round 3. The choices for player $i$ that survive round 3 are precisely the choices he can rationally make if he expresses 1-fold and 2-fold belief in rationality.

### 8.4.6 Common Belief in Rationality

Above we have outlined the first three rounds of an elimination procedure, and we have argued that (i) the choices that survive round 2 are precisely the choices that can rationally be made under 1-fold belief in rationality, and (ii) the choices that survive round 3 are precisely the choices that can rationally be made under 1-fold and 2-fold belief in rationality. If we continue in round 4 and further by the same steps, we arrive at the following recursive elimination procedure.

**Definition 8.4.1 (Iterated elimination of choices and second-order expectations)** *At the beginning, set up the decision problems for both players.*

***Round 1.*** *Eliminate for both players the choices that are strictly dominated in their decision problem. This yields the one-fold reduced decision problems.*

***Round 2.*** *Consider player $i$'s one-fold reduced decision problem. For every opponent's choice $c_j$, let $B_j^2(c_j)$ be the set of first-order beliefs $b_j$ for player $j$ such that there is a second-order expectation $e_j \in E_j$ for player $j$ that has the first-order belief $b_j$ and for which the choice $c_j$ is optimal. Let $E_i^2$ be the set of second-order expectations $e_i$ where*

$$e_i(\cdot \mid c_j) \text{ is in } B_j^2(c_j)$$

*for all choices $c_j$ to which $e_i$ assigns positive probability, and where $e_i(\cdot \mid c_j)$ is the first-order belief for player $j$ given by*

$$e_i(c_i \mid c_j) := \frac{e_i(c_j, c_i)}{\sum_{c_i' \in C_i} e_i(c_j, c_i')} \text{ for every choice } c_i \in C_i.$$

*Eliminate all choices for player $i$ that are not optimal for any second-order expectation in $E_i^2$. Do the same for player $j$. This yields the two-fold reduced decision problems.*

***Round 3.*** *Consider player $i$'s two-fold reduced decision problem. For every opponent's choice $c_j$, let $B_j^3(c_j)$ be the set of first-order beliefs $b_j$ for player $j$ such that there is a second-order expectation $e_j \in E_j^2$ for player $j$ that has the first-order belief $b_j$ and for which the choice $c_j$ is optimal. Let $E_i^3$ be the set of second-order expectations $e_i$ where*

$$e_i(\cdot \mid c_j) \text{ is in } B_j^3(c_j)$$

*for all choices $c_j$ to which $e_i$ assigns positive probability. Eliminate all choices for player $i$ that are not optimal for any second-order expectation in $E_i^3$. Do the same for player $j$.*

*And so on.*

*The choices that survive all elimination rounds are said to* ***survive*** *the* ***iterated elimination of choices and second-order expectations****.*

Note that at a given round $k$, the set $B_j^k(c_j)$ may be empty for an opponent's choice $c_j$. This happens precisely when the choice $c_j$ cannot be supported by any second-order expectation in $E_j^{k-1}$, and hence this choice $c_j$ has been eliminated in the previous round. In that case, it follows by the definition of the procedure that every second-order expectation $e_i \in E_i^k$ should assign probability zero to this opponent's choice $c_j$. Indeed, suppose that $e_i$ would assign a positive probability to $c_j$. Then, by definition, $e_i(\cdot \mid c_j)$ should be in $B_j^k(c_j)$, which cannot be since $B_j^k(c_j)$ is empty. As such, at every round, every second-order expectation that survives that round should assign probability zero to all opponent's choices that have been eliminated so far.

Recall that we argued above that round 2 of the procedure characterizes precisely those choices that can rationally be made if the player expresses 1-fold belief in rationality, and that round 3 of the procedure characterizes precisely those choices that can rationally be made if the player expresses up to 2-fold belief in rationality. By extending these arguments above to round 4 and further we conclude that, for every round $k \geq 2$, the choices that survive round $k$ of this procedure are precisely the choices that can rationally be made if the player expresses up to $(k-1)$-fold belief in rationality. As a consequence, the choices that survive all rounds are exactly the choices that can rationally be made under common belief in rationality. We thus obtain the following result.

**Theorem 8.4.1 (Procedure for common belief in rationality)** *(a) For every $k \in \{1, 2, 3, ...\}$, the choices that player $i$ can rationally make while expressing up to $k$-fold belief in rationality are precisely the choices that survive the first $k+1$ rounds of the iterated elimination of choices and second-order expectations.*

*(b) The choices that player $i$ can rationally make under common belief in rationality are exactly the choices that survive all rounds of the iterated elimination of choices and second-order expectations.*

In earlier chapters we have only seen elimination procedures that terminate within finitely many rounds. That is, for all of these procedures there is some round $k$ such that after this round no further choices, states and beliefs could be eliminated. This, however, is not true for the *iterated elimination of choices and second-order expectations.* In the next subsection we will see an example where in *every* round we can reduce the set of second-order expectations for both players compared to the previous round. In spite of this, it can still be shown that for every player there is at least one choice that survives all rounds of the procedure.

**Theorem 8.4.2 (At least one choice survives procedure)** *For every player there is at least one choice that survives all rounds of the iterated elimination of choices and second-order expectations.*

| **You** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 0 | 0 | 0 | 5 |
| *white* | 2 | 2 | 2 | 2 |

| **Barbara** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 2 | 2 | 2 | 2 |
| *white* | 5 | 0 | 0 | 0 |

Table 8.4.5 Decision problems for "Dinner with a strong preference for surprise"

Thus, even if the set of second-order expectations may decrease with every further round, we can always be sure that for every player at least one choice survives all of these rounds.

### 8.4.7 Examples

In this subsection we will illustrate the procedure by means of two examples. In the first, the procedure will terminate after five rounds, whereas in the second the procedure will keep restricting the set of second-order expectations forever, and will thus not terminate within finitely many rounds.

**Example 8.4: Dinner with a strong preference for surprise.**

A few weeks after "The black and white dinner with a twist", Chris invites you and Barbara once again for a *black and white dinner*. Your preferences have changed compared to last time: If you believe, with probability 1, that Barbara wears *white* and believe, with probability 1, that Barbara believes, with probability 1, that you wear *white,* then you *strongly* prefer wearing *black* to wearing *white.* Recall that before, you only slightly preferred *black* to *white* under these circumstances. Similarly for Barbara. These new conditional preferences can be modelled by the decision problems in Table 8.4.5.

Which color(s) can you rationally wear under common belief in rationality? To answer this question we apply the *iterated elimination of choices and second-order expectations.*

**Round 1.** As no choice for you or Barbara is strictly dominated in the decision problems above, we cannot eliminate any choice in round 1.

**Round 2.** We determine, for every choice $c_2$ of Barbara, the set of first-order beliefs $B_2^2(c_2)$ that support the choice $c_2$. To this purpose we first visualize the conditional preference relations for you and Barbara in Figure 8.4.4. Note that you prefer *black* to *white* when the probability you assign to the state $(w,w)$ is at least 0.4, and you prefer *white* to *black* otherwise. Please verify this. This yields the graphical representation of your conditional preference relation in the left-hand panel of Figure 8.4.4. Similarly for Barbara.

Hence, the second-order expectations that support Barbara's choice *white* are those that lie above the grey triangle in the right-hand panel. All of these second-order expectations induce a first-order belief that assigns a probability of at least 0.4 to you choosing *black.* Hence, the set of first-order beliefs that support Barbara's choice *white* is

$$B_2^2(white) = \{b_2 \in \Delta(C_1) \mid b_2(black) \geq 0.4\}. \tag{8.4.9}$$

On the other hand, the second-order expectations that support Barbara's choice *black* lie below the grey triangle in the right-hand panel of that figure. It may be verified that every first-order belief for Barbara is induced by at least one such second-order expectation. As such, the set of first-order beliefs that support Barbara's choice *black* is

$$B_2^2(black) = \Delta(C_1).$$

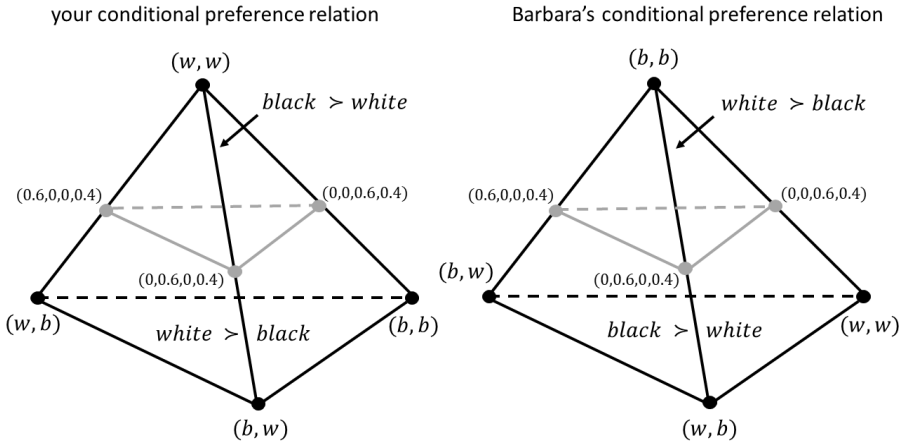



Figure 8.4.4 Conditional preference relations for you and Barbara in "Dinner with a strong preference for surprise"

The second-order expectations for you that are consistent with one-fold belief in rationality are thus given by the set

$$E_1^2 = \{e_1 \in E_1 \mid e_1(\cdot \mid white_2) \in B_2^2(white) \text{ and } e_1(\cdot \mid black_2) \in B_2^2(black)\}, \tag{8.4.10}$$

where the subindex 2 in $white_2$ and $black_2$ indicates that these choices belong to player 2 (Barbara). Since $B_2^2(black) = \Delta(C_1)$, the condition $e_1(\cdot \mid black_2) \in B_2^2(black)$ imposes no additional restriction on the second-order expectation $e_1$. By (8.4.9) and (8.4.10) we thus conclude that

$$E_1^2 = \{e_1 \in E_1 \mid e_1(black_1 \mid white_2) \geq 0.4\}.$$

This set is visualized in the left-hand panel of Figure 8.4.5. The set $E_1^2$ of second-order expectations for you that are consistent with one-fold belief in rationality are the ones below the dotted grey triangle. It can be seen that both choices for you are optimal for at least one second-order expectation in $E_1^2$, and hence we cannot eliminate any choice for you.

By symmetry, it can be verified that Barbara's second-order expectations which are consistent with one-fold belief in rationality are given by the set

$$E_2^2 = \{e_2 \in E_2 \mid e_2(white_2 \mid black_1) \geq 0.4\}.$$

Please verify this. This set can be found in the right-hand panel of Figure 8.4.5, below the dotted grey triangle. Also for Barbara, both choices are optimal for at least one second-order expectation in $E_2^2$, and hence we cannot eliminate any choice for Barbara either.

**Round 3.** We determine, for every choice $c_2$ of Barbara, the set $B_2^3(c_2)$ of first-order beliefs for Barbara that support the choice $c_2$ and that are consistent with one-fold belief in rationality. From the right-hand panel in Figure 8.4.5 it can be seen that the second-order expectations for Barbara in $E_2^2$ that support her choice *white* are the ones below the dotted grey triangle and above the solid grey triangle. What first-order beliefs do these second-order expectations induce?

Let us denote these second-order expectations by $E_2^2(white)$ – the second-order expectations in $E_2^2$ that support Barbara's choice *white*. From the figure, it can be seen that the extreme points of

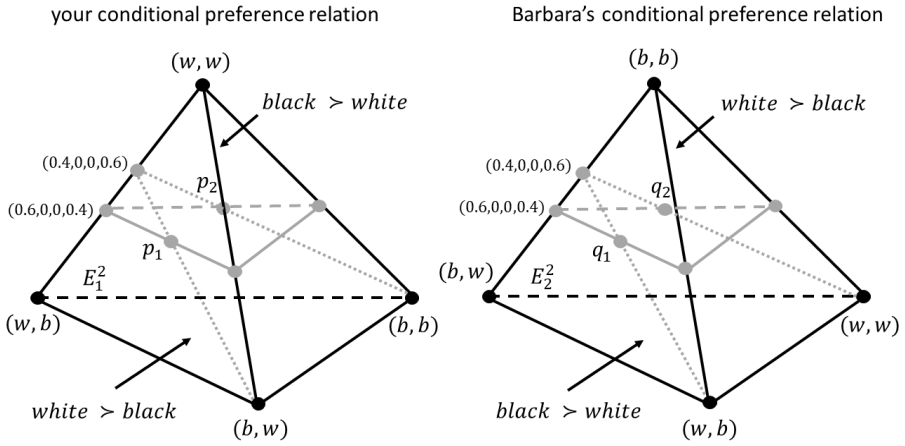



Figure 8.4.5 Second-order expectations consistent with one-fold belief in rationality in "Dinner with a strong preference for surprise"

the set $E_2^2(white)$ are $(0.6, 0, 0, 0.4)$, $(0.4, 0, 0, 0.6)$, $q_1$ and $q_2$. Recall that the first, second, third and fourth coordinate refer to the states $(b, w), (w, b), (w, w)$ and $(b, b)$, respectively. To compute $q_1$, note that $q_1$ is on the line between $(0, 1, 0, 0)$ and $(0.4, 0, 0, 0.6)$ and that the fourth coordinate is 0.4. That is, there is some $\lambda \in [0, 1]$ with

$$q_1 = (1 - \lambda) \cdot (0, 1, 0, 0) + \lambda \cdot (0.4, 0, 0, 0.6).$$

As the fourth coordinate must be 0.4, we obtain that

$$\lambda \cdot (0.6) = 0.4,$$

which yields $\lambda = 2/3$. Hence, we have that

$$q_1 = 1/3 \cdot (0, 1, 0, 0) + 2/3 \cdot (0.4, 0, 0, 0.6) = (4/15, 5/15, 0, 6/15).$$

In a similar way, it can be verified that

$$q_2 = (4/15, 0, 5/15, 6/15).$$

Note that the first and second extreme second-order expectation in $E_2^2(white)$, which are $(0.6, 0, 0, 0.4)$ and $(0.4, 0, 0, 0.6)$, induce the first-order belief that assigns probability 1 to your choice *black,* since the first and fourth coordinate refer to the states $(b, w)$ and $(b, b)$, respectively. Moreover, the other two extreme second-order expectations in $E_2^2(white)$, which are $q_1$ and $q_2$., induce the first-order belief that assigns probability $4/15 + 6/15 = 2/3$ to the choice *black.* As such, all second-order expectations in $E_2^2(white)$ induce first-order beliefs for which the probability assigned to your choice *black* lies between $2/3$ and 1.

By definition, $B_2^3(white)$ are those first-order beliefs for Barbara that are induced by some second-order expectation in $E_2^2(white)$. By our insight above, we thus conclude that

$$B_2^3(white) = \{b_2 \in \Delta(C_1) \mid b_2(black_1) \geq 2/3\}. \tag{8.4.11}$$

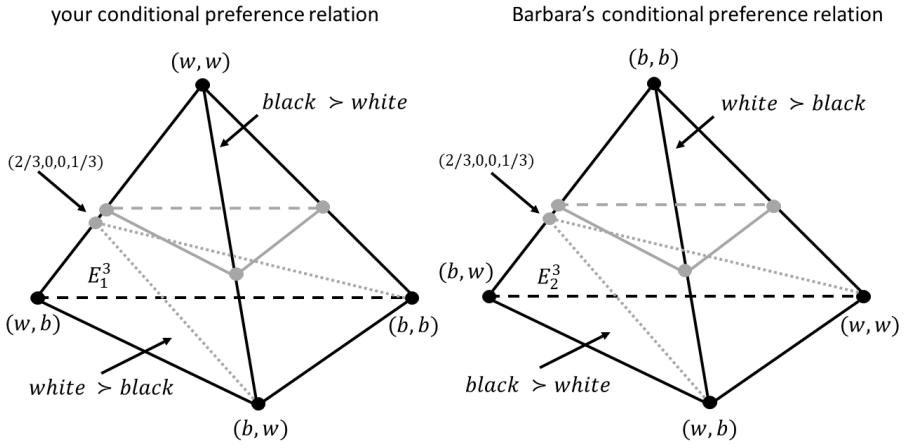



Figure 8.4.6 Second-order expectations consistent with one-fold and two-fold belief in rationality in "Dinner with a strong preference for surprise"

To determine $B_2^3(black)$ we look at the right-hand panel of Figure 8.4.5. Note that for Barbara, the set $E_2^2$ of second-order expectations consistent with one-fold belief in rationality contains every second-order expectation of the form $(1-\lambda)\cdot(b,w)+\lambda\cdot(w,b)$, where $\lambda \in [0,1]$. Since Barbara's choice *black* is optimal for every such second-order expectation, and every first-order belief for Barbara is induced by one such second-order expectation, we conclude that

$$B_2^3(black) = \Delta(C_1). \tag{8.4.12}$$

That is, every first-order belief for Barbara can be extended to a second-order expectation in $E_2^2$ for which Barbara's choice *black* is optimal.

The set $E_1^3$ of second-order expectations for you that are consistent with one-fold and two-fold belief in rationality is given by

$$E_1^3 = \{e_1 \in E_1 \mid e_1(\cdot \mid white_2) \in B_2^3(white) \text{ and } e_1(\cdot \mid black_2) \in B_2^3(black)\}. \tag{8.4.13}$$

Note that the condition $e_1(\cdot \mid black_2) \in B_2^3(black)$ imposes no additional restriction on $e_1$, since $B_2^3(black) = \Delta(C_1)$ by (8.4.12). If we combine (8.4.13) and (8.4.11) we thus get that

$$E_1^3 = \{e_1 \in E_1 \mid e_1(black_1 \mid white_2) \geq 2/3\}.$$

Similarly to what we have seen in round 2, we can visualize the set $E_1^3$ by the left-hand panel of Figure 8.4.6, where $E_1^3$ contains all the second-order expectations below the dotted grey triangle. Please verify this.

As this dotted grey triangle lies completely below the upper grey triangle, where you are indifferent between *white* and *black*, we conclude you prefer *white* to *black* for every second-order expectation in $E_1^3$. In particular, there is no second-order expectation in $E_1^3$ for which wearing *black* is optimal for you, and hence we can eliminate your choice *black* in round 3.

Similarly, it can be verified that the set $E_2^3$ of second-order expectations for Barbara that are consistent with one-fold and two-fold belief in rationality are those below the dotted grey triangle in the right-hand panel of Figure 8.4.6. As this dotted grey triangle lies completely below the upper grey

| **You** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *white* | 2 | 2 | 2 | 2 |

| **Barbara** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 2 | 2 | 2 | 2 |

Table 8.4.6 Reduced decision problems in round 3 of procedure for "Dinner with a strong preference for surprise"

triangle, where Barbara is indifferent between *white* and *black,* we see that Barbara prefers *black* to *white* for every second-order expectation in $E_2^3$. As such, there is no second-order expectation in $E_2^3$ for which wearing *white* is optimal for Barbara, and hence we can eliminate Barbara's choice *white* in round 3. We thus obtain the reduced decision problems for you and Barbara in Table 8.4.6.

**Round 4.** Note that Barbara's choice $white_2$ has been eliminated in Round 3, because it was not optimal for any second-order expectation $e_2 \in E_2^3$. As such, $B_2^4(white_2)$ is empty. But then, it follows that every second-order expectation $e_1 \in E_1^4$ must assign probability zero to Barbara's choice $white_2$. As such, we have that the set $E_1^4$ of second-order expectations that are consistent with up to three-fold belief in rationality is given by

$$E_1^4 = \{e_1 \in E_1 \mid e_1(w,b) = e_1(w,w) = 0 \text{ and } e_1(\cdot \mid black_2) \in B_2^4(black)\}. \tag{8.4.14}$$

Hence, we only need to determine $B_2^4(black)$ in order to obtain the set $E_1^4$. From the right-hand panel of Figure 8.4.6 it can be seen that every second-order expectation for Barbara of the form $e_2 = (1-\lambda)\cdot(b,w) + \lambda \cdot (w,b)$, where $\lambda \in [0,1]$, is in $E_2^3$ and supports Barbara's choice *black.* As every first-order belief for Barbara is induced by one such second-order expectation, we conclude that

$$B_2^4(black) = \Delta(C_1). \tag{8.4.15}$$

By combining (8.4.14) and (8.4.15) we obtain that

$$E_1^4 = \{e_1 \in E_1 \mid e_1(w,b) = e_1(w,w) = 0\}.$$

This set $E_1^4$ has been visualized as the dotted grey line in the left-hand panel of Figure 8.4.7.

Similarly, the set $E_2^4$ of second-order expectations for Barbara that are consistent with up to three-fold belief in rationality can be visualized by the grey dotted line in the right-panel of that figure.

**Round 5.** We first compute $E_1^5$ – the set of second-order expectations that are consistent with up to four-fold belief in rationality. From the left-hand panel of Figure 8.4.7 it can be seen that $E_1^4$ consists of the second-order expectations that assign probability 1 to Barbara's choice *black.* Hence, we obtain that

$$E_1^5 = \{e_1 \in E_1 \mid e_1(w,b) = e_1(w,w) = 0 \text{ and } e_1(\cdot \mid black_2) \in B_2^5(black)\}. \tag{8.4.16}$$

In order to derive $E_1^5$ we therefore only need to compute $B_2^5(black)$ – the set of first-order beliefs for Barbara that are induced by some second-order expectation $e_2 \in E_2^4$ and that support her choice *black.*

From the right-hand panel in Figure 8.4.7 it can be seen that all second-order expectations for Barbara in $E_2^4$ assign probability 1 to your choice *white.* Hence, every second-order expectations for Barbara that is in $E_2^4$ and supports her choice *black* has the first-order belief that assigns probability 1 to your choice *white.* This means, in turn, that

$$B_2^5(black) = \{b_2 \in \Delta(C_1) \mid b_2(white_1) = 1\}. \tag{8.4.17}$$

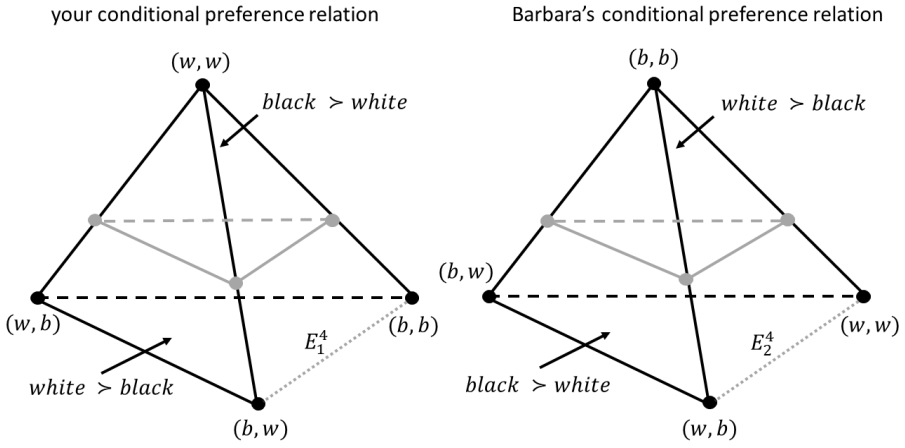



Figure 8.4.7 Second-order expectations consistent with up to three-fold belief in rationality for "Dinner with a strong preference for surprise"

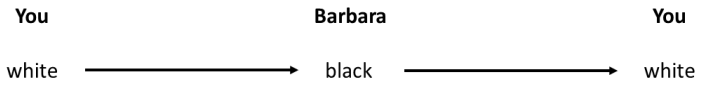



Figure 8.4.8 Beliefs diagram for "Dinner with a strong preference for surprise"

By combining (8.4.16) and (8.4.17) we then obtain that

$$E_1^5 = \{e_1 \in E_1 \mid e_1(w,b) = e_1(w,w) = 0 \text{ and } e_1(white_1 \mid black_2) = 1\},$$

which means that

$$E_1^5 = \{e_1 \in E_1 \mid e_1(b,w) = 1\}.$$

In a similar fashion, it can be shown that

$$E_2^5 = \{e_2 \in E_2 \mid e_2(w,b) = 1\}.$$

Hence, $E_1^5$ and $E_2^5$ only contain one second-order expectation, and this is where the procedure ends.

We thus conclude that under common belief in rationality, you can only rationally wear *white* and Barbara can only rationally wear *black.* However, note that in rounds 1 and 2 of the procedure we were not able to eliminate any choices. We could only eliminate the choice *black* for you and the choice *white* for Barbara in round 3, after we "sufficiently" reduced the set of second-order expectations for you and and Barbara. Indeed, in round 3 we saw that there was no surviving second-order expectation for you for which it is optimal to choose *black,* whereas in rounds 1 and 2 there still *were* surviving second-order expectations for which wearing *black* was optimal. Similarly for Barbara.

To see why your choice *white* and Barbara's choice *black* can rationally be made under common belief in rationality, consider the beliefs diagram in Figure 8.4.8. Similarly to the beliefs diagram in Figure 8.4.3 it may be verified that the unique belief hierarchy for you and the unique belief hierarchy

| **You** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 0 | 0 | 0 | 8 |
| *white* | 2 | 2 | 2 | 2 |

| **Barbara** | $(b,b)$ | $(b,w)$ | $(w,b)$ | $(w,w)$ |
|---|---|---|---|---|
| *black* | 2 | 2 | 2 | 2 |
| *white* | 8 | 0 | 0 | 0 |

Table 8.4.7 Decision problems for "Dinner with a huge preference for surprise"

for Barbara in this figure express common belief in rationality. As under this belief hierarchy it is optimal for you to wear *white,* we conclude that you can indeed rationally wear *white* under common belief in rationality. Similarly, Barbara's choice *black* is optimal for her belief hierarchy in this figure, and hence we see that Barbara can rationally choose *black* under common belief in rationality. Note, however, that under common belief in rationality you are not able to surprise Barbara by the color you wear, nor is it possible for Barbara to surprise you.

We will next discuss an example where the procedure does not terminate after finitely many rounds. That is, with every round we are able to decrease the sets of second-order expectations, and this reduction process never stops.

***Example 8.5: Dinner with a huge preference for surprise.**

Chris really enjoys the black and white dinners with you and Barbara, and he has therefore invited you and Barbara once again for a dinner at his house. Compared to the example "Dinner with a strong preference for suprise" there is only one change in your conditional preference relation: If you believe, with probability 1, that Barbara wears *white*, and you believe, with probability 1, that Barbara believes, with probability 1, that you wear *white,* the intensity by which you prefer *black* to *white* is even stronger than before. And similarly for Barbara. Suppose that the new conditional preference relations for you and Barbara are given by Table 8.4.7.

Which color(s) can you rationally wear under common belief in rationality? To answer this question we will apply the procedure *iterated elimination of choices and second-order expectations.* As we will see, the procedure does not terminate within finitely many rounds, since in every round we are able to eliminate some second-order expectations for both players.

**Round 1.** We cannot eliminate any choice in round 1 since no choice is strictly dominated for you or for Barbara.

**Round 2.** We first determine, for every choice $c_2$ of Barbara, the set of first-order beliefs $B_2^2(c_2)$ that support the choice $c_2$. To do this, we start by depicting the conditional preference relations for you and Barbara in Figure 8.4.9. Indeed, you prefer *black* to *white* when the probability you assign to the state $(w,w)$ is at least 0.25, and you prefer *white* to *black* otherwise. Please verify this. This yields the left-hand panel of Figure 8.4.9. Similarly for Barbara.

Therefore, the second-order expectations that support Barbara's choice *white* are those that lie above the grey triangle in the right-hand panel. These second-order expectations all induce a first-order belief that assigns a probability of at least 0.25 to you choosing *black.* Hence, the set of first-order beliefs that support Barbara's choice *white* is

$$B_2^2(white) = \{b_2 \in \Delta(C_1) \mid b_2(black) \geq 0.25\}. \tag{8.4.18}$$

At the same time, the second-order expectations that support Barbara's choice *black* lie below the grey triangle in the right-hand panel of that figure. It may be checked that every first-order belief for

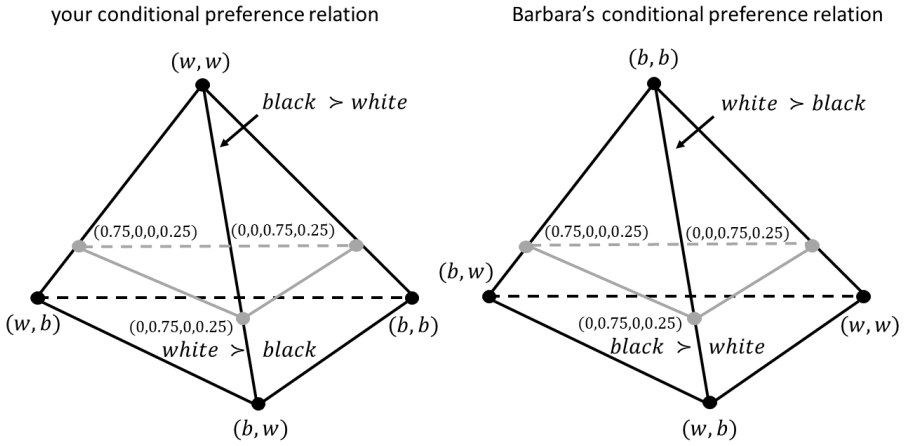



Figure 8.4.9 Conditional preference relations for 'Dinner with a huge preference for surprise"

Barbara is induced by at least one such second-order expectation. Consequently, the set of first-order beliefs that support Barbara's choice *black* is

$$B_2^2(black) = \Delta(C_1).$$

As such, the second-order expectations for you that are consistent with one-fold belief in rationality are given by

$$E_1^2 = \{e_1 \in E_1 \mid e_1(\cdot \mid white_2) \in B_2^2(white) \text{ and } e_1(\cdot \mid black_2) \in B_2^2(black)\}. \tag{8.4.19}$$

As $B_2^2(black) = \Delta(C_1)$, the condition $e_1(\cdot \mid black_2) \in B_2^2(black)$ imposes no additional restriction on the second-order expectation $e_1$. By (8.4.18) and (8.4.19) we thus conclude that

$$E_1^2 = \{e_1 \in E_1 \mid e_1(black_1 \mid white_2) \geq 0.25\}.$$

The set $E_1^2$ has been depicted in the left-hand panel of Figure 8.4.10. More precisely, the set $E_1^2$ of second-order expectations for you that are consistent with one-fold belief in rationality are the points below the dotted grey triangle. It can be seen that both choices for you are optimal for at least one second-order expectation in $E_1^2$, and hence we cannot eliminate any choice for you.

Similarly, it can be verified that Barbara's second-order expectations which are consistent with one-fold belief in rationality are given by the set

$$E_2^2 = \{e_2 \in E_2 \mid e_2(white_2 \mid black_1) \geq 0.25\}.$$

This set can be found in the right-hand panel of Figure 8.4.10, below the dotted grey triangle. Also for Barbara, both choices are optimal for at least one second-order expectation in $E_2^2$, and hence we cannot eliminate any choice for Barbara either.

**Round 3.** We first determine, for every choice $c_2$ of Barbara, the set $B_2^3(c_2)$ of first-order beliefs for Barbara that support the choice $c_2$ and that are consistent with one-fold belief in rationality. From the right-hand panel in Figure 8.4.10 it can be seen that the second-order expectations for Barbara in $E_2^2$ that support her choice *white* are the ones below the dotted grey triangle and above the solid grey triangle. What first-order beliefs do these second-order expectations have?

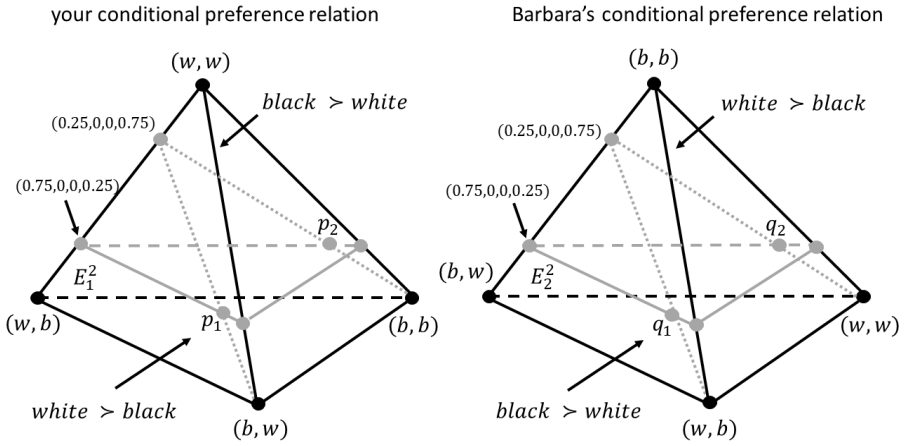



Figure 8.4.10 Second-order expectations consistent with one-fold belief in rationality for "Dinner with a huge preference for surprise"

We denote these second-order expectations by $E_2^2(white)$ – the second-order expectations in $E_2^2$ that support Barbara's choice *white.* From Figure 8.4.10 it can be seen that the extreme points of the set $E_2^2(white)$ are $(0.75, 0, 0, 0.25)$, $(0.25, 0, 0, 0.75)$, $q_1$ and $q_2$. We now compute $q_1$. Observe that $q_1$ is on the line between $(0, 1, 0, 0)$ and $(0.25, 0, 0, 0.75)$ and that the fourth coordinate is 0.25. That is, there is some $\lambda \in [0, 1]$ with

$$q_1 = (1 - \lambda) \cdot (0, 1, 0, 0) + \lambda \cdot (0.25, 0, 0, 0.75).$$

As the fourth coordinate must be 0.25, we obtain that

$$\lambda \cdot (0.75) = 0.25,$$

which yields $\lambda = 1/3$. Hence, we have that

$$q_1 = 2/3 \cdot (0, 1, 0, 0) + 1/3 \cdot (0.25, 0, 0, 0.75) = (1/12, 8/12, 0, 3/12).$$

In a similar way it can be checked that

$$q_2 = (1/12, 0, 8/12, 3/12).$$

Observe that the first and second extreme second-order expectation in $E_2^2(white)$, which are $(0.75, 0, 0, 0.25)$ and $(0.25, 0, 0, 0.75)$, induce the first-order belief that assigns probability 1 to your choice *black,* since the first and fourth coordinate refer to the states $(b, w)$ and $(b, b)$, respectively. Moreover, the other two extreme second-order expectations in $E_2^2(white)$, which are $q_1$ and $q_2$, induce the first-order belief that assigns probability $1/12 + 3/12 = 1/3$ to the choice *black.* Therefore, all second-order expectations in $E_2^2(white)$ induce first-order beliefs for which the probability assigned to your choice *black* lies between 1/3 and 1.

By definition, $B_2^3(white)$ are those first-order beliefs for Barbara that are induced by some second-order expectation in $E_2^2(white)$. Based on our insight above we thus conclude that

$$B_2^3(white) = \{b_2 \in \Delta(C_1) \mid b_2(black_1) \geq 1/3\}. \tag{8.4.20}$$

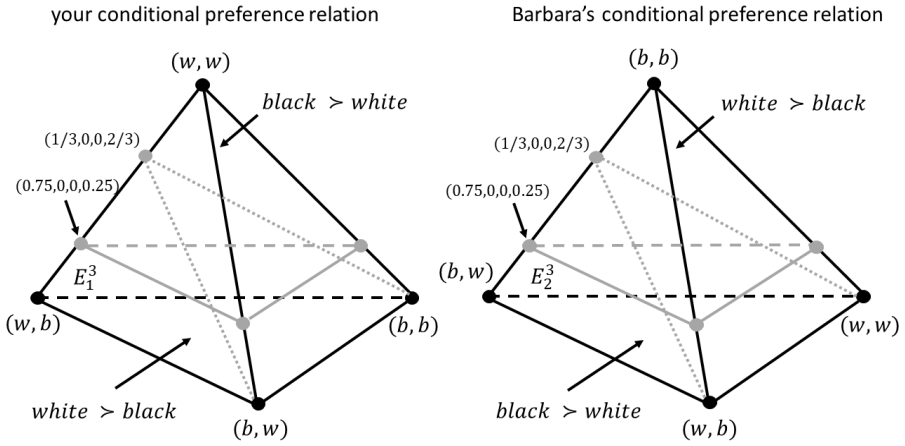



Figure 8.4.11 Second-order expectations consistent with up to two-fold belief in rationality for "Dinner with a huge preference for surprise"

We next determine $B_2^3(black)$. For this, we concentrate on the right-hand panel of Figure 8.4.10. Note that for Barbara, the set $E_2^2$ of second-order expectations consistent with one-fold belief in rationality contains every second-order expectation of the form $(1-\lambda)\cdot(b,w)+\lambda\cdot(w,b)$, where $\lambda\in[0,1]$. Since Barbara's choice *black* is optimal for every such second-order expectation, and every first-order belief for Barbara is induced by one such second-order expectation, we conclude that

$$B_2^3(black)=\Delta(C_1). \tag{8.4.21}$$

Thus, every first-order belief for Barbara can be extended to a second-order expectation in $E_2^2$ for which Barbara's choice *black* is optimal.

By definition, the set $E_1^3$ of second-order expectations for you that are consistent with one-fold and two-fold belief in rationality is given by

$$E_1^3=\{e_1\in E_1 \mid e_1(\cdot \mid white_2)\in B_2^3(white) \text{ and } e_1(\cdot \mid black_2)\in B_2^3(black)\}. \tag{8.4.22}$$

The condition $e_1(\cdot \mid black_2)\in B_2^3(black)$ imposes no additional restriction on $e_1$, since $B_2^3(black)=\Delta(C_1)$ by (8.4.21). If we combine (8.4.22) and (8.4.20) we thus get that

$$E_1^3=\{e_1\in E_1 \mid e_1(black_1 \mid white_2)\geq 1/3\}.$$

Similarly to round 2, we can depict the set $E_1^3$ by the left-hand panel of Figure 8.4.11, where $E_1^3$ contains all the second-order expectations below the dotted grey triangle.

As for you, both of your choices are optimal for at least one second-order expectation in $E_1^3$, we cannot eliminate any choice for you. Similarly for Barbara.

In the rounds that follow, the set $E_1^k$ of second-order expectations for you that are consistent with up to $(k-1)$-fold belief in rationality becomes smaller and smaller, but it never completely disappears below the solid grey triangle in Figure 8.4.11. More precisely, we will show that for every round $k$ the set $E_1^k$ is given by the points below the dotted grey triangle in Figure 8.4.12, and similarly for $E_2^k$. We will show this by induction on $k$, starting with $k=2$.

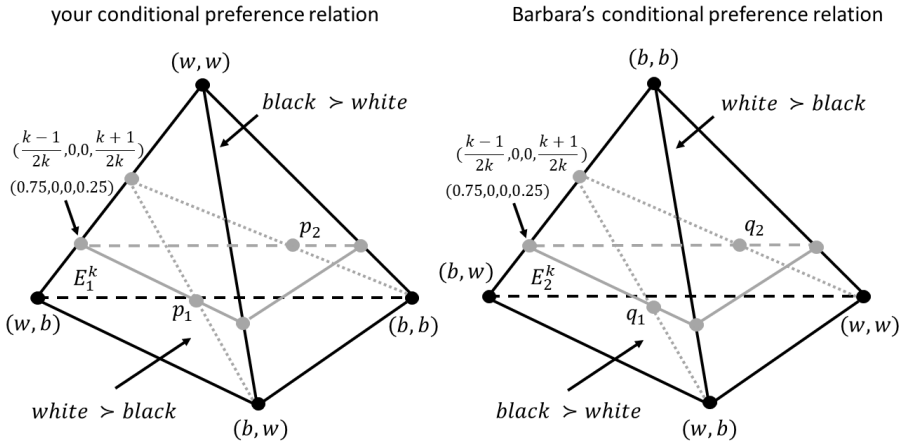



Figure 8.4.12 Second-order expectations consistent with up to $k-1$-fold belief in rationality for "Dinner with a huge preference for surprise"

For $k = 2$ we have seen that the second-order expectations for you that are consistent with 1-fold belief in rationality are the points below the dotted grey triangle in Figure 8.4.10. As

$$(0.25, 0, 0, 0.75) = (\tfrac{k-1}{2k}, 0, 0, \tfrac{k+1}{2k})$$

for $k = 2$, these are exactly the points below the dotted grey triangle in Figure 8.4.12. Similarly for Barbara.

Now, suppose that $k \geq 2$, and that $E_1^k$ and $E_2^k$ are the points below the dotted grey triangles in Figure 8.4.12. In other words, the extreme points of $E_2^k$ are

$$(\tfrac{k-1}{2k}, 0, 0, \tfrac{k+1}{2k}),\ (0, 1, 0, 0) \text{ and } (0, 0, 1, 0).$$

This will enable us to compute the set $E_1^{k+1}$ of second-order expectations for you that are consistent with up to $k$-fold belief in rationality.

Similarly to what we have done above, let $E_2^k(white)$ denote the set of the second-order expectations in $E_2^k$ that support Barbara's choice *white*. From Figure 8.4.12 it can be seen that the extreme points of the set $E_2^k(white)$ are $(\frac{k-1}{2k}, 0, 0, \frac{k+1}{2k})$, $(0.75, 0, 0, 0.25)$, $q_1$ and $q_2$. We now compute $q_1$. Observe that $q_1$ is on the line between $(0, 1, 0, 0)$ and $(\frac{k-1}{2k}, 0, 0, \frac{k+1}{2k})$ and that the fourth coordinate is 0.25. That is, there is some $\lambda \in [0, 1]$ with

$$q_1 = (1 - \lambda) \cdot (0, 1, 0, 0) + \lambda \cdot (\tfrac{k-1}{2k}, 0, 0, \tfrac{k+1}{2k}).$$

As the fourth coordinate must be 0.25, we obtain that

$$\lambda \cdot \tfrac{k+1}{2k} = 0.25,$$

which yields $\lambda = \frac{2k}{4k+4}$. Hence, we have that

$$q_1 = \tfrac{2k+4}{4k+4} \cdot (0, 1, 0, 0) + \tfrac{2k}{4k+4} \cdot (\tfrac{k-1}{2k}, 0, 0, \tfrac{k+1}{2k}) = (\tfrac{k-1}{4k+4}, \tfrac{2k+4}{4k+4}, 0, \tfrac{k+1}{4k+4}).$$

In a similar way, it can be checked that

$$q_2 = (\tfrac{k-1}{4k+4}, 0, \tfrac{2k+4}{4k+4}, \tfrac{k+1}{4k+4}).$$

Observe that the first and second extreme second-order expectation in $E_2^k(white)$, which are $(\frac{k-1}{2k}, 0, 0, \frac{k+1}{2k})$ and $(0.75, 0, 0, 0.25)$, induce the first-order belief that assigns probability 1 to your choice *black,* since the first and fourth coordinate refer to the states $(b, w)$ and $(b, b)$, respectively. Moreover, the other two extreme second-order expectations in $E_2^k(white)$, which are $q_1$ and $q_2$, induce the first-order belief that assigns probability

$$\frac{k-1}{4k+4} + \frac{k+1}{4k+4} = \frac{2k}{4k+4} = \frac{k}{2k+2}$$

to the choice *black.* Therefore, all second-order expectations in $E_2^2(white)$ induce first-order beliefs for which the probability assigned to your choice *black* lies between $\frac{k}{2k+2}$ and 1.

By definition, $B_2^{k+1}(white)$ are those first-order beliefs for Barbara that are induced by some second-order expectation in $E_2^k(white)$. Based on our insight above, we thus conclude that

$$B_2^{k+1}(white) = \{b_2 \in \Delta(C_1) \mid b_2(black_1) \geq \tfrac{k}{2k+2}\}. \tag{8.4.23}$$

We next determine $B_2^{k+1}(black)$. For this, we concentrate on the right-hand panel of Figure 8.4.12. Note that for Barbara, the set $E_2^k$ of second-order expectations consistent with up to $(k-1)$-fold belief in rationality contains every second-order expectation of the form $(1-\lambda) \cdot (b, w) + \lambda \cdot (w, b)$, where $\lambda \in [0, 1]$. Since Barbara's choice *black* is optimal for every such second-order expectation, and every first-order belief for Barbara is induced by one such second-order expectation, we conclude that

$$B_2^{k+1}(black) = \Delta(C_1). \tag{8.4.24}$$

Thus, every first-order belief for Barbara can be extended to a second-order expectation in $E_2^k$ for which Barbara's choice *black* is optimal.

By definition, the set $E_1^{k+1}$ of second-order expectations for you that are consistent with up to $k$-fold belief in rationality is given by

$$E_1^{k+1} = \{e_1 \in E_1 \mid e_1(\cdot \mid white_2) \in B_2^{k+1}(white) \text{ and } e_1(\cdot \mid black_2) \in B_2^{k+1}(black)\}. \tag{8.4.25}$$

The condition $e_1(\cdot \mid black_2) \in B_2^{k+1}(black)$ imposes no additional restriction on $e_1$, since $B_2^{k+1}(black) = \Delta(C_1)$ by (8.4.24). If we combine (8.4.25) and (8.4.23) we thus get that

$$E_1^{k+1} = \{e_1 \in E_1 \mid e_1(black_1 \mid white_2) \geq \tfrac{k}{2k+2}\}.$$

As

$$\frac{k}{2k+2} = \frac{(k+1)-1}{2(k+1)}$$

it follows that $E_1^{k+1}$ is the set of points below the dotted grey triangle in Figure 8.4.12 if we substitute $k+1$ for $k$. Similarly for $E_2^{k+1}$.

By induction on $k$ it thus follows that for every round $k$, the sets $E_1^k$ and $E_2^k$ contain the points below the two grey dotted triangles in Figure 8.4.12. Note that $\frac{k-1}{2k} \leq 0.75$ for every $k$, which implies that no choice can ever be eliminated in the procedure. As the sets $E_1^k$ and $E_2^k$ change with every round $k$, we conclude that the procedure does not terminate within finitely many rounds.

But the procedure still specifies which second-order expectations, and which choices, are possible for you and Barbara under *common* belief in rationality. Recall that for every $k$, the set of second-order expectations for you and Barbara that are consistent with up to $(k-1)$-fold belief in rationality are given by the points below the grey dotted triangles in Figure 8.4.12. If we let $k$ go to infinity, then the point $(\frac{k-1}{2k}, 0, 0, \frac{k+1}{2k})$ tends to $(1/2, 0, 0, 1/2)$. As such, the set of second-order expectations $E_1^*$ for

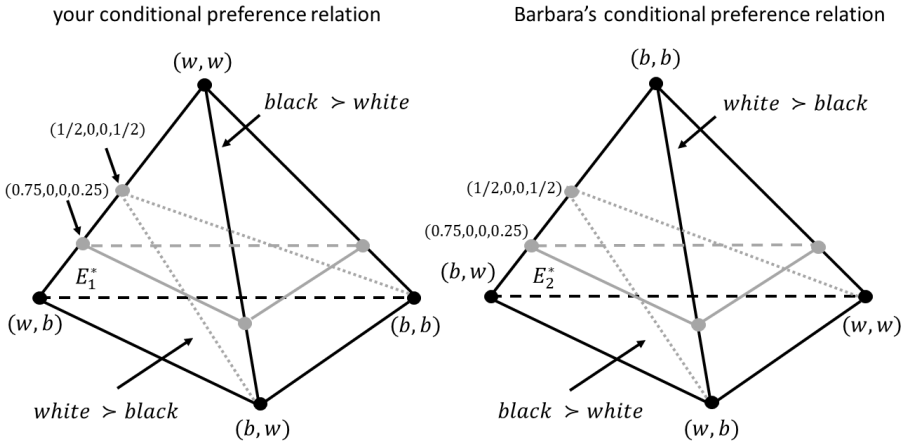



Figure 8.4.13 Second-order expectations consistent with common belief in rationality for "Dinner with a huge preference for surprise"

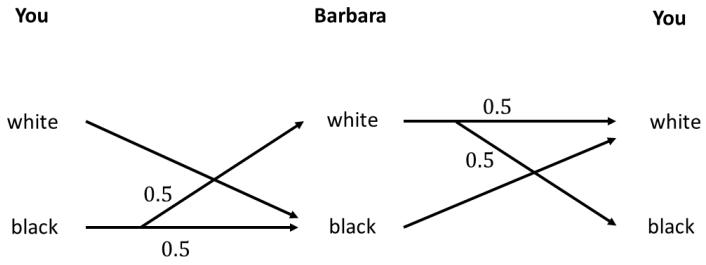



Figure 8.4.14 Beliefs diagram for "Dinner with a huge preference for surprise"

you that are consistent with *common* belief in rationality corresponds to the set of points below the grey dotted triangle in Figure 8.4.13, and similarly for Barbara.

Note that for you, there are second-order expectations consistent with common belief in rationality for which it is optimal to choose *white,* and others for which it is optimal to choose *black.* Hence, under common belief in rationality you can rationally choose *white* and *black,* but this cannot be concluded after finitely many steps of the procedure. Indeed, since the procedure does not terminate within finitely many steps, we must follow all the – infinitely many – steps of the procedure to conclude that both choices can rationally be made under common belief in rationality.

By symmetry, the same holds for Barbara. Hence, under common belief in rationality, both you and Barbara can rationally wear *white* and *black.* This conclusion is confirmed by the beliefs diagram in Figure 8.4.14. To verify the validity of the beliefs diagram, let us first focus on the second-order expectation we obtain for you if we start at your choice *white.* The second-order expectation is that you believe, with probability 1, that Barbara wears *black* and that Barbara believes that you wear *white.* Under this second-order expectation it is indeed optimal for you to wear *white.*

Next, consider the second-order expectation we obtain for you if we start at your choice *black.* You assign probability 0.5 to the event that "Barbara wears *white* and assigns probability 0.5 to you

wearing *white* and *black,* respectively", and you assign probability 0.5 to the event that "Barbara wears *black* and assigns probability 1 to you wearing *white*". The induced second-order expectation is thus given by

$$(0.25) \cdot (w, w) + (0.25) \cdot (w, b) + (0.5) \cdot (b, w).$$

Under this second-order expectation the expected utility of choosing *black* is $(0.25) \cdot 8 + (0.25) \cdot 0 + (0.5) \cdot 0 = 2$, whereas the expected utility of choosing *white* is also 2. Hence, you are indifferent between choosing *black* and *white* which means, in particular, that wearing *black* is optimal for you given this second-order expectation.

In a similar fashion, it can be shown that *white* is optimal for Barbara under the second-order expectation that starts at her choice *white,* and *black* is optimal for Barbara under the second-order expectation that starts at her choice *black.* As such, all solid arrows in the beliefs diagram are justified, which implies that all the belief hierarchies in this beliefs diagram express common belief in rationality. Hence, we can conclude that under common belief in rationality you can rationally wear *white* and *black,* and the same applies to Barbara.

### 8.4.8 Common Belief in Rationality is Always Possible

A natural question is: Can we always find, for every player, a belief hierarchy that expresses common belief in rationality? In other words, is it always possible to reason in accordance with common belief in rationality? The answer to this question is "yes", and it follows from the Theorems 8.4.1 and 8.4.2.

Indeed, Theorem 8.4.2 guarantees that for every player there is at least one choice that survives the iterated elimination of choices and second-order expectations, whereas Theorem 8.4.1 states that for every such choice we can always construct a type within an epistemic model that expresses common belief in rationality and supports that choice. But we can say a little more: In the proof of Theorem 8.4.1 it is shown that we can always construct a *single* epistemic model, with infinitely many types for both players, such that for every choice that survives the procedure there is a type within this particular epistemic model that expresses common belief in rationality and supports that choice. Moreover, this epistemic model has the property that *all* types which are present express common belief in rationality. We thus obtain the following result.

**Theorem 8.4.3 (Common belief in rationality is always possible)** *For every psychological game with finitely many choices we can construct an epistemic model, possibly with infinitely many types, in which all types express common belief in rationality.*

Compared to previous chapters there is an important difference here: For standard games, games with incomplete information, and games with unawareness, we could always build an epistemic model with *finitely* many types such that for every choice surviving the procedure there is a type that expresses common belief in rationality and supports this choice. For psychological games we need an epistemic model with *infinitely* many types to serve this purpose, at least if we use the construction in the proof of Theorem 8.4.1.

But suppose now that our objective is more modest: Instead of constructing an epistemic model that supports *every* choice surviving the procedure by a type that expresses common belief in rationality, we look for an epistemic model that contains for every player at least one type that expresses common belief in rationality. Can this be achieved by an epistemic model that contains finitely many types? The answer is "yes". In fact, it will follow from the existence of *simple* belief hierarchies that express common belief in rationality in Chapter 9. Based on that result, we are even able to construct an epistemic model with only *one* type for every player, such that this type induces a simple belief

hierarchy and expresses common belief in rationality. Hence, in Theorem 8.4.3, the phrase "possibly with infinitely many types" can be replaced by "with finitely many types".

### 8.4.9 Using the Procedure to Construct Epistemic Models

The procedure *iterated elimination of choices and second-order expectations* does not only deliver the choices that can rationally be made under common belief in rationality, but it also yields the second-order expectations that are consistent with common belief in rationality. As an illustration, consider the example "'Dinner with a huge preference for surprise". The second-order expectations that survive the procedure are given by Figure 8.4.13, and these are precisely the second-order expectations that are induced by belief hierarchies that express common belief in rationality.

Take the second-order expectation $e_1^*$ for you given by

$$e_1^* = (0.6) \cdot (w, b) + (0.4) \cdot (w, w).$$

From Figure 8.4.13 it is clear that the second-order expectation $e_1$ survives the procedure, and supports your choice *black*. Hence, there must be a belief hierarchy for you that induces the second-order expectation $e_1$ and expresses common belief in rationality. But how do we construct such a belief hierarchy? Below we present a method for finding such a belief hierarchy.

The method critically uses the sets of second-order expectations for you and Barbara that survive the procedure, and the properties they have. For both players $i$, let $E_i^*$ be the set of second-order expectations that survive the procedure. Moreover, for every choice $c_i$ that survives the procedure, let $B_i^*(c_i)$ be the set of first-order beliefs $b_i$ for which there is a second-order expectation $e_i \in E_i^*$ that has the first-order belief $b_i$ and for which the choice $c_i$ is optimal. By construction of the procedure we then have, for every second-order expectation $e_i \in E_i^*$, that

$$e_i(\cdot \mid c_j) \in B_j^*(c_j)$$

for every choice $c_j$ to which $e_i$ assigns positive probability. Or, equivalently, for every second-order expectation $e_i \in E_i^*$, and every choice $c_j$ to which $e_i$ assigns positive probability,

$$\begin{gathered} \text{there is } e_j \in E_j^* \text{ such that } e_j \text{ has the first-order belief } e_i(\cdot \mid c_j) \text{ and} \\ c_j \text{ is optimal for } e_j. \end{gathered} \tag{8.4.26}$$

Property (8.4.26) will be crucial for constructing a belief hierarchy for you that expresses common belief in rationality, and that has the second-order expectation $e_1^*$ above.

**Step 1.** We start by graphically depicting the second-order expectation $e_1^*$ in a beliefs diagram, in Figure 8.4.15.

**Step 2.** Since $e_1^* \in E_1^*$, it follows from (8.4.26) that there is some second-order expectation $e_2 \in E_2^*$ for Barbara that has the first-order belief

$$e_1^*(\cdot \mid \mathit{white}_2) = (0.6) \cdot \ \mathit{black}_1 + (0.4) \cdot \ \mathit{white}_1$$

and for which Barbara's choice $\mathit{white}_2$ is optimal. Here, we use the subindices 1 and 2 to indicate the player to which the choice belongs. Recall that you are player 1 and Barbara is player 2. From Figure 8.4.13 we see that we can choose, for instance, the second-order expectation

$$e_2 = (0.3) \cdot (b, w) + (0.3) \cdot (b, b) + (0.4) \cdot (w, b).$$

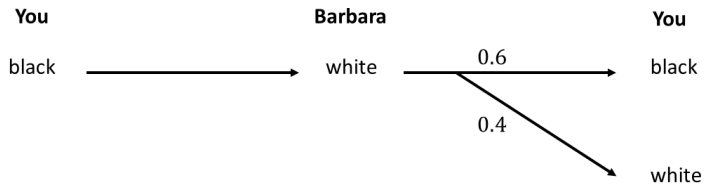



Figure 8.4.15 Constructing a belief hierarchy that expresses common belief in rationality: Step 1

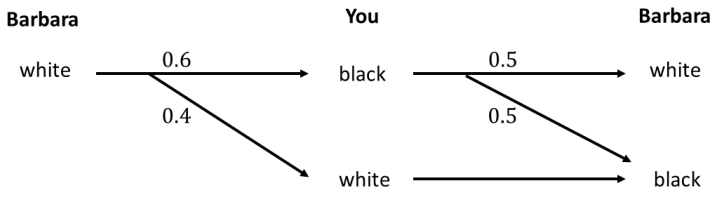



Figure 8.4.16 Constructing a belief hierarchy that expresses common belief in rationality: Step 2

This second-order expectation can be depicted graphically by the beliefs diagram in Figure 8.4.16.

**Step 3.** Since Barbara's second-order expectation $e_2$ is in $E_2^*$, we know from (8.4.26) that there is a second-order expectation $e_1[black_1]$ for you in $E_1^*$ that induces the first-order belief

$$e_2(\cdot \mid black_1) = (0.5) \cdot \ white_2 + (0.5) \cdot \ black_2$$

and for which your choice $black_1$ is optimal. Similarly, there is a second-order expectation $e_1[white_1]$ for you in $E_1^*$ that induces the first-order belief

$$e_2(\cdot \mid white_1) = \ black_1$$

and for which your choice $white_1$ is optimal. By looking at Figure 8.4.13 we see that can we choose the second-order expectations

$$e_1[black_1] = (0.25) \cdot (w, b) + (0.25) \cdot (w, w) + (0.5) \cdot (b, w)$$

and

$$e_1[white_1] = (b, w).$$

These two second-order expectations are graphically depicted in the beliefs diagram of Figure 8.4.17. Here, $e_1[black_1]$ is the second-order expectation that starts at your choice *black,* whereas $e_1[white_1]$ is the second-order expectation that starts at your choice *white.*

**Step 4.** Since $e_1[black_1]$ is in $E_1^*$, it follows from (8.4.26) that there is some second-order expectation $e_2[white_2]$ for Barbara that has the first-order belief

$$e_1[black_1](\cdot \mid white_2) = (0.5) \cdot \ black_1 + (0.5) \cdot \ white_1$$

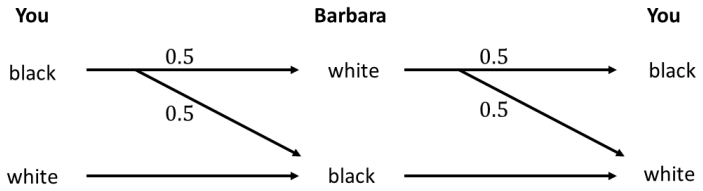



Figure 8.4.17 Constructing a belief hierarchy that expresses common belief in rationality: Step 3

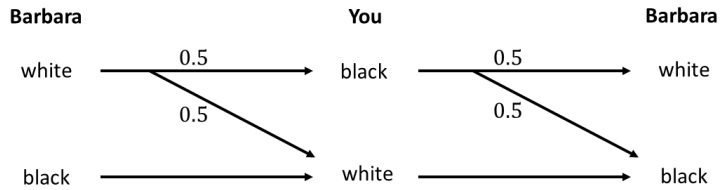



Figure 8.4.18 Constructing a belief hierarchy that expresses common belief in rationality: Step 4

and for which $white_2$ is optimal, and there is some second-order expectation $e_2[black_2]$ for Barbara that has the first-order belief

$$e_1[black_1](\cdot \mid black_2) = white_1$$

and for which $black_2$ is optimal. Similarly, since $e_1[white_1]$ is in $E_1^*$, we know from (8.4.26) that there is some second-order expectation $\hat{e}_2[black_2]$ for Barbara that has the first-order belief

$$e_1[white_1](\cdot \mid black_2) = white_1$$

and for which $black_2$ is optimal. From Figure 8.4.13 we see that we can choose

$$e_2[white_2] = (0.25) \cdot (b, w) + (0.25) \cdot (b, b) + (0.5) \cdot (w, b)$$

and

$$e_2[black_2] = \hat{e}_2[black_2] = (w, b).$$

These two second-order expectations are graphically represented by the beliefs diagram in Figure 8.4.18.

**Step 5.** Since Barbara's second-order expectation $e_2[white_2]$ is in $E_2^*$, we know from (8.4.26) that there is a second-order expectation $\hat{e}_1[black_1]$ for you in $E_1^*$ that induces the first-order belief

$$e_2[white_2](\cdot \mid black_1) = (0.5) \cdot \ white_2 + (0.5) \cdot \ black_2$$

and for which your choice *black* is optimal. Moreover, there is a second-order expectation $\hat{e}_1[white_1]$ for you in $E_1^*$ that induces the first-order belief

$$e_2[white_2](\cdot \mid white_1) = \ black_2$$

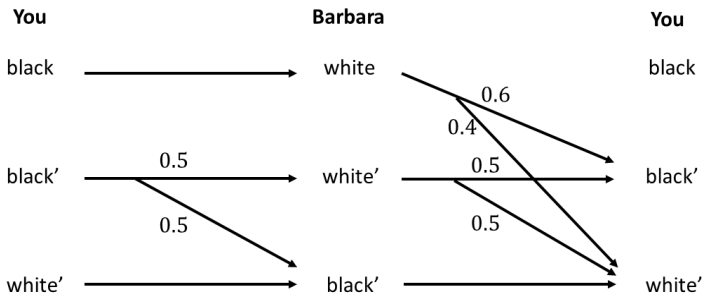



Figure 8.4.19 Constructing a belief hierarchy that expresses common belief in rationality: Pasting together the beliefs diagrams

and for which your choice *white* is optimal.

Similarly, since Barbara's second-order expectation $e_2[black_2]$ is in $E_2^*$, we know from (8.4.26) that there is a second-order expectation $\tilde{e}_1[white_1]$ for you in $E_1^*$ that induces the first-order belief

$$e_2[black_2](\cdot \mid white_1) = \ black_2$$

and for which your choice *white* is optimal.

If we compare this to Step 3, we see that we can choose

$$\hat{e}_1[black_1] = e_1[black_1] = (0.25) \cdot (w, b) + (0.25) \cdot (w, w) + (0.5) \cdot (b, w),$$

and we can choose

$$\hat{e}_1[white_1] = \tilde{e}_1[white_1] = e_1[white_1] = (b, w).$$

That is, in Step 5 we can choose exactly the same second-order expectations for you as in Step 3. But this means that from Step 5 onwards, we keep repeating Steps 3 and 4 forever. This results in the second-order expectations from Figures 8.4.17 and 8.4.18.

If we paste the beliefs diagrams from Figures 8.4.15, 8.4.16, 8.4.17 and 8.4.18 together, we obtain the larger beliefs diagram from Figure 8.4.19. Note that we need copies *black'* and *white'* of the choices for both players, since we need two different second-order expectations for you that support your choice *black,* and we need two different second-order expectations for Barbara that support her choice *white.*

**Question 8.4.3** *Translate the beliefs diagram from Figure 8.4.19 into an epistemic model.*

It may be verified that the belief hierarchy that starts at your choice *black* expresses common belief in rationality and has the second-order expectation

$$e_1^* = (0.6) \cdot (w, b) + (0.4) \cdot (w, w),$$

as required.

This method can be generalized to every psychological game: If we take an arbitrary psychological game, any choice $c_i$ that survives the procedure, and any second-order expectation $e_i$ that supports $c_i$ and survives the procedure, we can always construct, in the way described above, a belief hierarchy for player $i$ that supports the choice $c_i$, holds the second-order expectation $e_i$, and expresses common belief in rationality.

In fact, this is precisely the way we proceed in the proof of Theorem 8.4.1: For every choice $c_i$ that survives the procedure we take some second-order expectation $e_i[c_i]$ that survives the procedure and supports the choice $c_i$. By relying on property (8.4.26) we then construct, by following the method above, a type $t_i[c_i]$ that expresses common belief in rationality, holds the second-order expectation $e_i$, and supports the choice $c_i$.

## 8.5 States-First Procedure

In this section we present a variation on the procedure of *iterated elimination of choices and second-order expectations* which we call the *states-first procedure.* In the new procedure we start by recursively eliminating choices and states as long as we can, resulting in possibly reduced decision problems. We subsequently apply the iterated elimination of choices and second-order expectations to these reduced decision problems. The practical advantage of this new procedure is that we start with the "easy" elimination of choices and states, and reserve the more tedious steps in the iterated elimination of choices and second-order eliminations until later, when the decision problems – hopefully – have already been reduced. We argue that the alternative procedure yields exactly the same output as the original procedure, which implies that the states-first procedure can also be used to find those choices that can rationally be made under common belief in rationality. We finally illustrate the states-first procedure by a new example.

### 8.5.1 Elimination of States

Recall that every step of the procedure *iterated elimination of choices and second-order expectations* consists of reducing the set of second-order expectations, and subsequently eliminating those choices that are no longer optimal for any second-order expectation that survives. Especially the reduction of the set of second-order expectations may be a difficult step in many examples.

Sometimes, this reduction may be mimicked by the elimination of states, in which case the reduction becomes a lot easier to perform. Suppose, for instance, that we are at round $k$, and that the choice $c_j$ for player $j$ has been eliminated in the previous round. Then, as we have seen, every second-order expectation in $E_i^k$ must assign probability zero to this choice $c_j$. That is, at round $k$ we can safely eliminate for player $i$ all states that involve opponent's choices that have been eliminated at the previous round, since second-order expectations for player $i$ at round $k$ will assign probability zero to such states.

Now assume that player $i$'s own choice $c_i$ has been eliminated in the previous round $k-1$. What consequences does this have for player $i$'s second-order expectations at further rounds? By the same argument as above, every second-order expectation for opponent $j$ in $E_j^k$ must assign probability zero to this choice $c_i$. Hence, for every opponent's choice $c_j$, every first-order belief in $B_j^{k+1}(c_j)$ must assign probability zero to the choice $c_i$. Now, take any second-order expectation $e_i$ for player $i$ in $E_i^{k+1}$. Then,

for every opponent's choice $c_j$ to which $e_i$ assigns a positive probability it must hold that

$$e_i(\cdot \mid c_j) \text{ is in } B_j^{k+1}(c_j).$$

In particular, we see that for every choice $c_j$ to which $e_i$ assigns a positive probability, the conditional probability distribution $e_i(\cdot \mid c_j)$ on $C_i$ must assign probability zero to the choice $c_i$. But then, the second-order expectation $e_i$ must assign probability zero to all states that involve player $i$'s choice $c_i$. Summarizing, we see that if the choice $c_i$ is eliminated at round $k-1$, then every second-order expectation for player $i$ in round $k+1$ must assign probability zero to $c_i$. This can be mimicked by removing, at round $k+1$, all states in $i$'s decision problem that involve the choice $c_i$. In fact, we could already remove these states at round $k$, since they will anyhow be eliminated in the round that follows.

Altogether, we see that if a choice is eliminated at a particular round, then in the next round we can eliminate all states that contain this choice. Such eliminations thus constitute the "easy" reductions of the sets of second-order expectations in the procedure.

What would happen if we would perform all these "easy" reductions at the beginning of our procedure, and only afterwards start with the more tedious reductions of the sets of second-order expectations? Would we end up with the same output? The answer is "yes". The reason is that, similarly to the other procedures considered so far in this book, also the *iterated elimination of choices and second-order expectations* is order independent. That is, if at some rounds we do not eliminate some choices or second-order expectations that could have been eliminated, then we are still guaranteed to end up with the same end result, provided we do not forget to perform these eliminations forever. In particular, if during the first rounds we only do the "easy" reductions described above, and only start with the more tedious reductions later, then we will get the same output as under the original procedure.

This alternative procedure, which we call the *states-first* procedure, may have some practical advantages over the original procedure, especially when there are many choices involved. If we start with the "easy" reductions first, by iteratedly eliminating choices and states from the decision problems, then we hopefully end up with heavily reduced decision problems, with much less choices and states than before. This, in turn, makes it a lot easier to perform the more tedious reductions, since we start with much smaller decision problems. To formally define the *states-first* procedure, we must first explain precisely what we mean by the iterated elimination of choices and states mentioned above.

### 8.5.2 Iterated Elimination of Choices and States

Remember from above that, if a choice gets eliminated at a certain round, then in the next round we can eliminate all states that involve this choice. What we are actually doing by such an elimination of states is to exclude all second-order expectations that assign positive probability to this state.

Now assume that we have eliminated some states, and that we consider all second-order expectations that only assign positive probability to the states that remain in the reduced decision problem so obtained. Then, by Theorem 2.6.1, the choices that are optimal for such a second-order expectation are precisely the choices that are not strictly dominated in the reduced decision problem. All these considerations give rise to the following procedure, which we call the *iterated elimination of choices and states.*

**Definition 8.5.1 (Iterated elimination of choices and states)** *At the beginning, set up the decision problems for both players.*

***Round 1.*** *Eliminate for both players the choices that are strictly dominated in their decision problem. This yields the one-fold reduced decision problems.*

***Round 2.*** *Consider player $i$'s one-fold reduced decision problem. Eliminate all states $(c_j, c_i)$ where either the choice $c_j$ or the choice $c_i$ has been eliminated at the previous round. In the reduced decision problem so obtained, eliminate all choices for player $i$ that are strictly dominated. Do the same for player $j$. This yields the two-fold reduced decision problem.*

***Round 3.*** *Consider player $i$'s two-fold reduced decision problem. Eliminate all states $(c_j, c_i)$ where either the choice $c_j$ or the choice $c_i$ has been eliminated at the previous round. In the reduced decision problem so obtained, eliminate all choices for player $i$ that are strictly dominated. Do the same for player $j$. This yields the three-fold reduced decision problem. And so on.*

*The choices that survive all elimination rounds are said to* ***survive*** *the* ***iterated elimination of choices and states****.*

Hence, this procedure amounts to performing, in a recursive fashion, all the "easy" reductions in the *iterated elimination of choices and second-order expectations.*

### 8.5.3 States-First Procedure

As already mentioned above, the *states-first* procedure is obtained if we first perform the *iterated elimination of choices and states* until we can go no further. We then obtain a (possibly) reduced decision problem for the two players. Starting from these reduced decision problems we then apply the *iterated elimination of choices and second-order expectations.*

**Definition 8.5.2 (States-First Procedure)** *At the beginning, set up the decision problems for both players.*

***Step 1.*** *Apply the iterated elimination of choices and states until no further eliminations are possible. This yields, for both players, a reduced decision problem.*

***Step 2.*** *Take these reduced decision problems as a starting point, and then apply the iterated elimination of choices and second-order expectations.*

*The choices and second-order expectations that remain at the end are said to survive the* ***states-first procedure.***

From a practical viewpoint, the *states-first* procedure may be more attractive than the original *iterated elimination of choices and second-order expectations.* The reason is that step 1, the "easy" step, is often able to drastically reduce the decision problems for both players. In that case, the "tedious" reduction steps, which are all gathered in step 2, will be easier to carry out since the number of states on which the second-order expectations operate is smaller than before. To illustrate this, consider the following example.

**Example 8.6: Exceeding Barbara's expectations.**

Yesterday, Barbara and you have won a spectacular prize at the national lottery: Within seven weeks from now, you have the opportunity to record a song at the famous Abbey Road Studios, and the revenue from the sales goes entirely to the two of you. Of course, you both need to practice before you are able to do a decent recording. The problem, however, is that you and Barbara are not good at singing, and therefore practicing comes at a considerable mental cost.

This afternoon you will gather at Chris' house, where you and Barbara must specify how many weeks you are willing to practice. To make things easy, suppose that you can only choose between

practicing for one week, three weeks, five weeks and seven weeks, and the same applies to Barbara. Assume that the total revenue from selling the record, in thousands of euros, is equal to $2 \cdot w_1 \cdot w_2$, where $w_1$ and $w_2$ are the numbers of weeks that you and Barbara will practice, respectively. Of course, you will split this revenue equally between Barbara and you, and hence your income from selling the record is $w_1 \cdot w_2$. Note that the additional income from practicing two more weeks is increasing in Barbara's effort: The more she practices, the more profitable it becomes to practice yourself. On the other hand, your mental costs of practicing for $w_1$ weeks (expressed in terms of thousands of euros) is $w_1^2$.

Finally, it is important for you not to disappoint Barbara by the number of weeks you are willing to practice. More precisely, if you are actually willing to practice for $w_1$ weeks, and Barbara believes, with probability 1, that you are willing to practice for $w_1'$ weeks, with $w_1 > w_1'$, then you receive a mental bonus (expressed in terms of thousands of euros) of $w_1 - w_1'$ for exceeding Barbara's expectation. The mental bonus is thus equal to the amount by which your choice is believed to exceed Barbara's expectation. If $w_1 \leq w_1'$, you receive no such mental bonus.

To illustrate how the (expected) mental bonus is computed, consider the situation where you are willing to practice for 3 weeks, and where you believe, with probability 1, that Barbara believes with probability 0.4 that you are willing to practice for 1 week and that she believes with probability 0.6 that you are willing to practice for 7 weeks. Then, this belief assigns probability 0.4 to the event that $w_1 = 3$ and $w_1' = 1$, and probability 0.6 to the event that $w_1 = 3$ and $w_1' = 7$. The expected mental bonus is thus $(0.4) \cdot (3 - 1) + (0.6) \cdot 0 = 0.8$.

Be careful: One could also argue that in the belief above, you believe that Barbara believes that, in expectation, you are willing to practice for $(0.4) \cdot 1 + (0.6) \cdot 7 = 4.6$ weeks. In view of this, one could be tempted to say that your mental bonus is 0, since the actual number of weeks you want to practice, which is 3, lies below the *expected* number of weeks Barbara believes you will practice, which is 4.6. However, this is wrong: The expected mental bonus is *not* based on the *expected* number of weeks Barbara believes you will practice, but rather on the *belief* that you have about the *definite* number of weeks Barbara believes you are willing to practice.

Your total expected utility is given by your income from selling the record, minus the mental costs from practicing, plus a possible mental bonus from exceeding Barbara's expectation. That is, if you choose to practice for $w_1$ weeks, Barbara chooses to practice for $w_2$ weeks, and Barbara believes that you will practice for $w_1'$ weeks, then your utility is

$$u_1(w_1, (w_2, w_1')) := \begin{cases} w_1 \cdot w_2 - w_1^2 + (w_1 - w_1'), & \text{if } w_1 > w_1' \\ w_1 \cdot w_2 - w_1^2, & \text{otherwise} \end{cases} . \tag{8.5.1}$$

Note that the pairs $(w_2, w_1')$ are precisely the states in your decision problem. The utilities for Barbara are given by a similar expression. That is, also Barbara$\cup$ cares about exceeding the other person's expectations.

**Question 8.5.1** *Consider the first- and second-order belief for you as specified by the partial beliefs diagram in Figure 8.5.1.*

*(a) Write down the second-order expectation $e_1$ that is induced by these beliefs.*

*(b) Suppose you are willing to practice for five weeks, and hold the beliefs as given by Figure 8.5.1. Using the formula in (8.5.1), calculate your expected utility.*

Note that the utilities in (8.5.1) give rise to a psychological game, since the preferences over your choices depend on what you believe that Barbara believes about the number of weeks you are willing

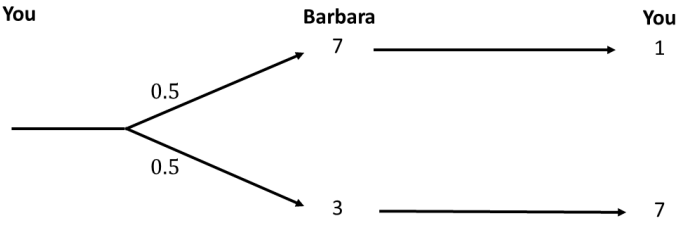



Figure 8.5.1 Partial beliefs diagram for "Exceeding Barbara's expectations"

| **You** | $(1,1)$ | $(1,3)$ | $(1,5)$ | $(1,7)$ | $(3,1)$ | $(3,3)$ | $(3,5)$ | $(3,7)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 2 |
| 3 | $-4$ | $-6$ | $-6$ | $-6$ | 2 | 0 | 0 | 0 |
| 5 | $-16$ | $-18$ | $-20$ | $-20$ | $-6$ | $-8$ | $-10$ | $-10$ |
| 7 | $-36$ | $-38$ | $-40$ | $-42$ | $-22$ | $-24$ | $-26$ | $-28$ |

| **You** | $(5,1)$ | $(5,3)$ | $(5,5)$ | $(5,7)$ | $(7,1)$ | $(7,3)$ | $(7,5)$ | $(7,7)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 4 | 4 | 4 | 6 | 6 | 6 | 6 |
| 3 | 8 | 6 | 6 | 6 | 14 | 12 | 12 | 12 |
| 5 | 4 | 2 | 0 | 0 | 14 | 12 | 10 | 10 |
| 7 | $-8$ | $-10$ | $-12$ | $-14$ | 6 | 4 | 2 | 0 |

Table 8.5.1 Decision problem for "Exceeding Barbara's expectations"

to practice, and similarly for Barbara. In other words, your preferences depend on your second-order beliefs, and the same applies to Barbara. The decision problem for you is given by Table 8.5.1. The decision problem for Barbara is similar.

The question is: Which choices can you rationally make under common belief in rationality? To answer this question, we will use the *states-first* procedure. We start by the iterated elimination of choices and states.

**Round 1.** In your decision problem from Table 8.5.1 your choice 7 is strictly dominated by your choice 5, and can thus be eliminated. Similarly for Barbara. This yields the one-fold decision problem for you in Table 8.5.2.

| **You** | $(1,1)$ | $(1,3)$ | $(1,5)$ | $(1,7)$ | $(3,1)$ | $(3,3)$ | $(3,5)$ | $(3,7)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 2 |
| 3 | $-4$ | $-6$ | $-6$ | $-6$ | 2 | 0 | 0 | 0 |
| 5 | $-16$ | $-18$ | $-20$ | $-20$ | $-6$ | $-8$ | $-10$ | $-10$ |

| **You** | $(5,1)$ | $(5,3)$ | $(5,5)$ | $(5,7)$ | $(7,1)$ | $(7,3)$ | $(7,5)$ | $(7,7)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 4 | 4 | 4 | 6 | 6 | 6 | 6 |
| 3 | 8 | 6 | 6 | 6 | 14 | 12 | 12 | 12 |
| 5 | 4 | 2 | 0 | 0 | 14 | 12 | 10 | 10 |

Table 8.5.2 One-fold reduced decision problem for "Exceeding Barbara's expectations"

| **You** | $(1,1)$ | $(1,3)$ | $(1,5)$ | $(3,1)$ | $(3,3)$ | $(3,5)$ | $(5,1)$ | $(5,3)$ | $(5,5)$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 2 | 2 | 2 | 4 | 4 | 4 |
| 3 | $-4$ | $-6$ | $-6$ | 2 | 0 | 0 | 8 | 6 | 6 |

Table 8.5.3 Two-fold reduced decision problem for "Exceeding Barbara's expectations"

| **You** | $(1,1)$ | $(1,3)$ | $(3,1)$ | $(3,3)$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 2 | 2 |
| 3 | $-4$ | $-6$ | 2 | 0 |

Table 8.5.4 Three-fold reduced decision problem for "Exceeding Barbara's expectations"

**Round 2.** From your decision problem in Table 8.5.2 we start by eliminating all states $(w_2, w_1')$ where either $w_2$ or $w_1'$ is equal to 7. In the reduced decision problem that remains, your choice 5 is strictly dominated by 3 and can thus be eliminated. Similarly for Barbara. This yields the two-fold reduced decision problem for you in Table 8.5.3.

**Round 3.** From the decision problem in Table 8.5.3 we start by eliminating all states $(w_2, w_1')$ where either $w_2$ or $w_1'$ is equal to 5. In the remaining decision problem no choice is strictly dominated, and hence the iterated elimination of choices and states terminates here. The final decision problem for the iterated elimination of choices and states is thus given by Table 8.5.4.

We now take the reduced decision problem from Table 8.5.4 as the input for applying the *iterated elimination of choices and second-order expectations.*

**Round 1.** In the decision problem from Table 8.5.4 no choice is strictly dominated. Hence, no choice can be eliminated for you, and similarly for Barbara.

**Round 2.** For both choices of Barbara, 1 and 3, we wish to find the sets $B_2^2(1)$ and $B_2^2(3)$ of first-order beliefs that support these choices. To do so, we first visualize Barbara's conditional preference relation in the right-hand panel of Figure 8.5.2. Note that choosing 3 is only optimal for Barbara if her second-order expectation assigns probability 1 to the state $(3,1)$. In that case, she would actually be indifferent between choosing 1 and 3. For all other second-other expectations Barbara would prefer 1 to 3.

As a consequence, the only first-order belief for Barbara that supports her choice 3 is the one that assigns probability 1 to you choosing 3. That is,

$$B_2^2(3) = \{3\}. \tag{8.5.2}$$

On the other hand, every first-order belief for Barbara can be extended to a second-order expectation for which her choice 1 is optimal. Hence,

$$B_2^2(1) = \Delta(C_1). \tag{8.5.3}$$

We are now ready to derive the second-order expectations for you in $E_1^2$. By definition, $E_1^2$ contains exactly those second-order expectations $e_1$ such that

$$e_1(\cdot \mid c_2) \text{ is in } B_2^2(c_2)$$

for every choice $c_2$ of Barbara to which $e_1$ assigns a positive probability. Together with (8.5.2) and (8.5.3) we thus see that $E_1^2$ contains precisely those second-order expectations $e_1$ such that

$$e_1(\cdot \mid 3_2) = 3_1 \tag{8.5.4}$$

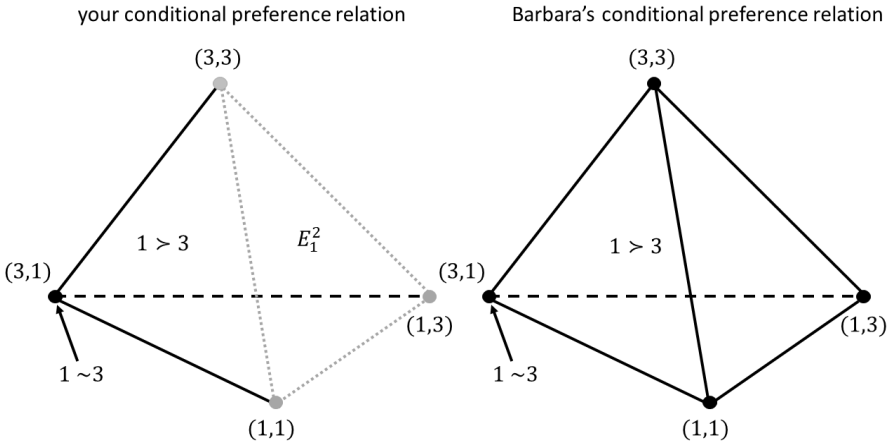



Figure 8.5.2 Second-order expectations for round 2 in "Exceeding Barbara's expectations"

whenever $e_1$ assigns positive probability to $3_2$, and

$$e_1(\cdot \mid 1_2) \in \Delta(C_1) \tag{8.5.5}$$

whenever $e_1$ assigns positive probability to $1_2$. Here, the subindices 1 and 2 indicate the player to which the choice belongs. As (8.5.5) is not a restriction, the set $E_1^2$ contains exactly those second-order expectations $e_1$ that satisfy (8.5.4). Visually, these are the second-order expectations in the grey dotted triangle of the left-hand panel in Figure 8.5.2.

Note that your conditional preference relation is the same as the one for Barbara, as can be seen in the left-hand panel of this figure. By comparing your conditional preference relation with the set $E_1^2$ of second-order expectations, we see that there is no second-order expectation in $E_1^2$ for which your choice 3 is optimal. We can thus eliminate your choice 3. Similarly for Barbara.

Since we are left with only one choice for both players, which is the choice 1, we conclude that under common belief in rationality you can only rationally practice for one week. In particular, it will not be optimal to try to exceed Barbara's expectations under common belief in rationality. To reach this conclusion, the *iterated elimination of choices and states* was not enough, as this procedure left the choices 1 and 3 for both players. To eliminate the choice 3, we had to apply the more tedious *iterated elimination of choices and second-order expectations* to the reduced decision problem in Table 8.5.4.

At the same time, applying the *iterated elimination of choices and second-order expectations* right from the beginning, to the full decision problem in Table 8.5.1, would have been extremely difficult and cumbersome, since we would have to deal with 16 states. This example thus shows that using the *states-first* procedure can be very convenient from a practical point of view.

To conclude this example, we show that you can indeed rationally choose to practice for 1 week under common belief in rationality. Consider the beliefs diagram in Figure 8.5.3. It can be verified that your unique belief hierarchy expresses common belief in rationality, and that your choice 1 is optimal for the second-order expectation $(1, 1)$ induced by this belief hierarchy. Therefore, you can indeed rationally practice for one week under common belief in rationality.

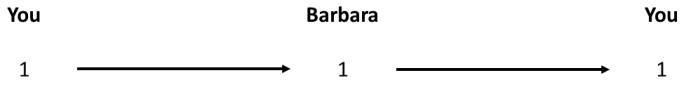



Figure 8.5.3 Beliefs diagram for "Exceeding Barbara's expectations"

# 8.6 When Elimination of Choices and States is Enough

In this section we will identify conditions under which the iterated elimination of choices and states is sufficient to capture all the choices that can rationally be made under common belief in rationality. This will be useful, since the iterated elimination of choices and states is an easy procedure to use – much easier than the *iterated elimination of choices and second-order expectations,* and it is therefore important to know when we are allowed to use this easier procedure. As a first step, we go back to the example "The black and white dinner", and try to see intuitively why the elimination of choices and states is not sufficiently fine-grained to yield the choices that are possible under common belief in rationality. The reason, as we will see, is that in this example, trying to support certain choices under common belief in rationality may induce conflicting conditions on higher-order beliefs. This insight will then help us to identify conditions where this problem of conflicting conditions cannot occur, and where the elimination of choices and states *is* sufficient to yield all the choices that can rationally be made under common belief in rationality.

## 8.6.1 Conflicting Conditions on Higher-Order Beliefs

In Section 8.4.2 we have seen that in order to obtain the choices you can rationally make under common belief in rationality, the recursive elimination of states and choices may not be enough. Recall the example "The black and white dinner" with the decision problems as given in Table 8.4.3. We saw that under common belief in rationality you cannot rationally choose *black,* but at the same time your choice *black* cannot be eliminated by the recursive elimination of states and choices alone. But what exactly causes this discrepancy?

The problem is that in order to justify your choice *black* while believing in Barbara's rationality, we obtain two conflicting conditions on your second-order belief. On the one hand, your choice *black* can only be optimal if your second-order expectation is $(w, w)$. In particular, your choice *black* can only be optimal if your second-order belief assigns probability 1 to the event that Barbara assigns probability 1 to you choosing *white.*

On the other hand, if your second-order expectation is $(w, w)$ then you must assign probability 1 to Barbara choosing *white.* Note that for Barbara, choosing *white* is only optimal if she holds the second-order expectation $(b, b)$. Hence, if you believe in Barbara's rationality, then your second-order belief must assign probability 1 to the event that Barbara assigns probability 1 to you choosing *black.*

Summarizing, we see that in order to support your choice *black,* your second-order belief must assign probability 1 to the event that Barbara assigns probability 1 to you choosing *white,* whereas the belief in Barbara's rationality implies that your second-order belief must assign probability 1 to the event that Barbara assigns probability 1 to you choosing *black.* Clearly, these two conditions are at odds, and therefore you cannot rationally choose *black* if you believe in Barbara's rationality.

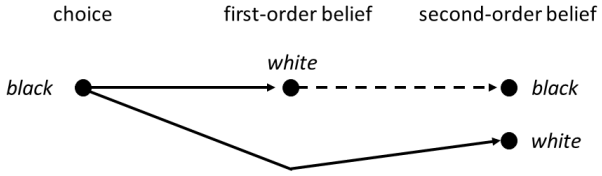



Figure 8.6.1 Causality diagram for "The black and white dinner"

The tension between these two conditions can be visualized by the *causality diagram* in Figure 8.6.1. Here, the restrictions on the first- and second-order belief that follow from supporting your choice *black* are represented by the two solid arrows. The restriction on the second-order belief induced by the belief in Barbara's rationality is given by the dashed arrow.

The problem is that the removal of choices and states alone is not sufficiently fine-grained to identify this tension. Indeed, in the example every choice is optimal for at least one second-order expectation, and hence no choice can be removed in the first round. As a consequence, no choice can be removed *at all* if we restrict to eliminating choices and states. What this procedure overlooks is the fact that the only second-order expectation which supports your choice *black* – the second-order expectation $(w, w)$ – is inconsistent with the belief in Barbara's rationality.

### 8.6.2 When Conflicting Conditions Cannot Arise

In the example above we have seen that in order to support a given choice by a second-order expectation, the belief in the opponent's rationality may yield contradictory conditions on the second-order belief. When this occurs, we conclude that the choice cannot rationally be made under common belief rationality. At the same time, the iterated elimination of choices and states is not sufficiently fine-grained to identify these contradictory conditions, and may therefore fail to eliminate this choice.

This raises the following question: Are there classes of psychological games where such conflicting conditions on higher-order beliefs can simply not arise? As we will see, the answer is yes, and these will precisely be the games where the iterated elimination of choices and states is sufficient to identify the choices that are possible under common belief in rationality. As a first illustration, consider the following example.

**Example 8.7: Barbara's birthday.**

Tomorrow it will be Barbara's birthday, and you want to buy her a nice gift. Suppose you can choose between buying her a *necklace,* a *ring,* or a *bracelet,* but that you prefer buying a *necklace* to buying a *ring,* and that you prefer buying a *ring* to buying a *bracelet.* At the same time you would like it if your gift would come as a surprise to her.

Barbara, on the other hand, would like to guess what you will be buying, and she prefers to guess correctly. Hence, she can choose between guessing you bought a *necklace,* a *ring,* or a *bracelet.* However, you do not care about Barbara's actual guess: You only care about the gift you buy, and about the degree by which you believe to surprise Barbara.

This situation can be modeled by the psychological game in Table 8.6.1. In your decision problem, $(\cdot, n)$ represents the collection of states where you believe that Barbara believes that you will buy a *necklace.* Similarly for the other two columns. The $\cdot$ indicates that your belief about Barbara's

| **You** | $(\cdot, n)$ | $(\cdot, r)$ | $(\cdot, b)$ |
|---|---|---|---|
| *necklace* | 0 | 3 | 3 |
| *ring* | 2 | 0 | 2 |
| *bracelet* | 1 | 1 | 0 |

| **Barbara** | $(n, \cdot)$ | $(r, \cdot)$ | $(b, \cdot)$ |
|---|---|---|---|
| *necklace* | 1 | 0 | 0 |
| *ring* | 0 | 1 | 0 |
| *bracelet* | 0 | 0 | 1 |

Table 8.6.1 Decision problems for "Barbara's birthday"

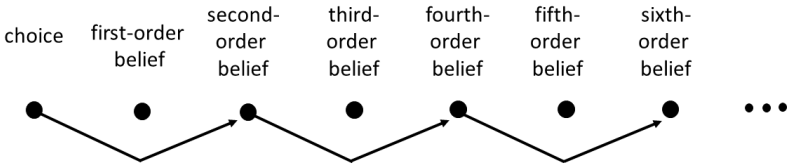



Figure 8.6.2 Causality diagram for "Barbara's birthday"

choice – that is, your first-order belief – does not matter for your preference relation over choices. However, it is relevant for your decision problem what Barbara believes about your choice. Hence, your preferences only depend on your second-order belief but not on your first-order belief.

Similarly, in Barbara's decision problem, $(n, \cdot)$ represents the collection of states where she believes that you will buy a *necklace.* The $\cdot$ thus means that Barbara's preferences do not depend on her second-order belief, only on her first-order belief.

Similarly to Figure 8.6.1 we can also make a causality diagram for this particular game, and see that no conflicting conditions on higher-order beliefs can arise if we want to support a choice under common belief in rationality. Consider the causality diagram in Figure 8.6.2. Indeed, if we wish to support a choice for you under common belief in rationality by a second-order expectation, then it is only relevant what we choose as the second-order belief, since your preferences do not depend on your first-order belief. In turn, to justify the second-order belief (what you believe that Barbara believes that you will buy) under common belief in rationality, it is only relevant what we choose as the fourth-order belief (what you believe that Barbara believes that you believe that Barbara believes that you will buy). The reason is similar: If you believe that Barbara believes that you buy the gift $g$ then, under common belief in rationality, you must believe that Barbara believes that you hold a second-order belief that makes buying the gift $g$ optimal. This, in turn, leads to a fourth-order belief. Your third-order belief is not relevant here. We could of course continue in this way: To justify your fourth-order belief it is only relevant what we choose as your sixth-order belief, for similar reasons, and so on. By continuing in this way we never get contradicting conditions on any of your higher-order beliefs.

Because of this, the *iterated elimination of choices and states,* as discussed in Section 8.5.2, will be sufficient to identify all choices you can rationally make under common belief in rationality. To see why, let us first apply this procedure.

**Round 1.** In your decision problem, the choice *bracelet* is strictly dominated by the randomized choice $(0.4)\cdot$ *necklace* + $(0.6)\cdot$ *ring,* and can thus be eliminated. This yields the one-fold reduced decision problems in Table 8.6.2.

| **You** | $(\cdot, n)$ | $(\cdot, r)$ | $(\cdot, b)$ |
|---|---|---|---|
| *necklace* | 0 | 3 | 3 |
| *ring* | 2 | 0 | 2 |

| **Barbara** | $(n, \cdot)$ | $(r, \cdot)$ | $(b, \cdot)$ |
|---|---|---|---|
| *necklace* | 1 | 0 | 0 |
| *ring* | 0 | 1 | 0 |
| *bracelet* | 0 | 0 | 1 |

Table 8.6.2 One-fold reduced decision problems in "Barbara's birthday"

| **You** | $(\cdot, n)$ | $(\cdot, r)$ |
|---|---|---|
| *necklace* | 0 | 3 |
| *ring* | 2 | 0 |

| **Barbara** | $(n, \cdot)$ | $(r, \cdot)$ |
|---|---|---|
| *necklace* | 1 | 0 |
| *ring* | 0 | 1 |

Table 8.6.3 Two-fold reduced decision problems in "Barbara's birthday"

**Round 2.** Since your choice *bracelet* has been eliminated at round 1, we can eliminate the states $(\cdot, b)$ in your decision problem, and we can eliminate the states $(b, \cdot)$ in Barbara's decision problem. Subsequently, in Barbara's reduced decision problem, her guess *bracelet* becomes strictly dominated by the randomized guess $(0.5)\cdot$ *necklace* + $(0.5)\cdot$ *ring,* and can thus be eliminated. This leads to the two-fold reduced decision problems in Table 8.6.3.

**Round 3.** Since Barbara's choice *bracelet* has been eliminated in Round 2, we can eliminate the states $(b, n)$ and $(b, r)$ in your decision problem, and we can eliminate the states $(n, b)$ and $(r, b)$ in Barbara's decision problem. This leads to the three-fold reduced decision problem in Table 8.6.4, after which the procedure of *iterated elimination of choices and states* ends.

Hence, your choices *necklace* and *ring* survive the *iterated elimination of choices and states.* We will see that under common belief in rationality you can indeed rationally make these two choices. To show this we proceed in two steps: We first design a partial beliefs diagram where every surviving choice for you is supported by a second-order belief that only points to surviving choices for you. We then complete the beliefs diagram by adding first-order beliefs where you believe in Barbara's rationality.

On the basis of your reduced decision problem in Table 8.6.4 we can design the partial beliefs diagram in Figure 8.6.3, where every surviving choice for you has been supported by a second-order belief that only points to surviving choices for you. Note that we have not specified your first-order beliefs yet, which is why we call it a partial beliefs diagram. This has been indicated by the dots, which refer to some unspecified choices for Barbara.

Subsequently, we can complete this partial beliefs diagram to a full beliefs diagram, by additionally specifying your first-order beliefs. Of course, we will do so in a way that reflects your belief in Barbara's rationality. Consider your choice *necklace,* which is supported by the second-order belief where you believe that Barbara believes that you will buy a *ring.* If, in addition, you believe in Barbara's rationality, you must necessarily believe that Barbara will guess that you buy a *ring.* As such, your

| **You** | $(n, n)$ | $(n, r)$ | $(r, n)$ | $(r, r)$ |
|---|---|---|---|---|
| *necklace* | 0 | 3 | 0 | 3 |
| *ring* | 2 | 0 | 2 | 0 |

| **Barbara** | $(n, n)$ | $(n, r)$ | $(r, n)$ | $(r, r)$ |
|---|---|---|---|---|
| *necklace* | 1 | 1 | 0 | 0 |
| *ring* | 0 | 0 | 1 | 1 |

Table 8.6.4 Three-fold reduced decision problems in "Barbara's birthday"

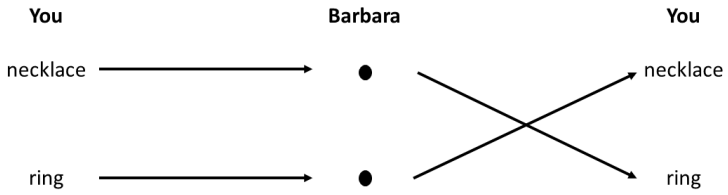



Figure 8.6.3 Partial beliefs diagram for "Barbara's birthday"

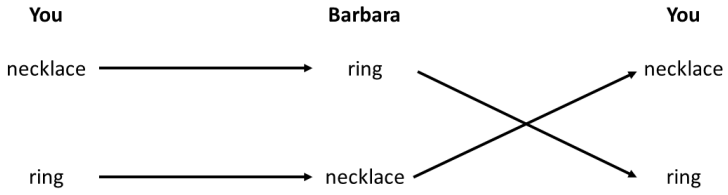



Figure 8.6.4 Full beliefs diagram for "Barbara's birthday"

choice *necklace* will be supported by the combination of a first- and second-order belief where you believe that Barbara guesses *ring,* and you believe that Barbara believes that you will buy a *ring.* See the full beliefs diagram in Figure 8.6.4. By a similar reasoning, your choice *ring* can be supported by the combination of a first- and second-order belief where you believe that Barbara guesses *necklace,* and you believe that Barbara believes that you will buy a *necklace.*

Since all the arrows are solid, we conclude on the basis of Figure 8.6.4 that you can rationally buy either a *necklace* or a *ring* under common belief in rationality. Therefore, the *iterated elimination of choices and states* delivers precisely the choices you can rationally make under common belief in rationality.

The method we have used above can be applied to any psychological game where player 1's preferences only depend on his second-order belief and player 2's preferences only depend on his first-order belief. As a consequence, for every such game the choices that player 1 can rationally make under common belief in rationality can be found by using the *iterated elimination of choices and states.* Note that we are interested in the choices that *player 1* can rationally make under common belief in rationality, and hence we are approaching the game from player 1's viewpoint.

We can go a bit further here: Suppose that player 1's preferences depend only on his second-order belief, whereas player 2's preferences depend either on his first- and second-order belief, or on his second-order belief alone. Then, we can apply a method similar to the one above to show that player 1's choices that can rationally be made under common belief in rationality are exactly given by the *iterated elimination of choices and states.*

This is even true if player 1's preferences only depend on his first-order belief and player 2's preferences depend only on his second-order belief. To see why, let us consider an example which is basically the same as "Barbara's birthday", but now with the roles of you and Barbara reversed.

| **You** | $(h,\cdot)$ | $(g,\cdot)$ | $(s,\cdot)$ |
|---|---|---|---|
| *hat* | 1 | 0 | 0 |
| *gloves* | 0 | 1 | 0 |
| *scarf* | 0 | 0 | 1 |

| **Barbara** | $(\cdot,h)$ | $(\cdot,g)$ | $(\cdot,s)$ |
|---|---|---|---|
| *hat* | 0 | 3 | 3 |
| *gloves* | 2 | 0 | 2 |
| *scarf* | 1 | 1 | 0 |

Table 8.6.5 Decision problems for "Your birthday"

| **You** | $(h,h)$ | $(h,g)$ | $(g,h)$ | $(g,g)$ |
|---|---|---|---|---|
| *hat* | 1 | 1 | 0 | 0 |
| *gloves* | 0 | 0 | 1 | 1 |

| **Barbara** | $(h,h)$ | $(h,g)$ | $(g,h)$ | $(g,g)$ |
|---|---|---|---|---|
| *hat* | 0 | 3 | 0 | 3 |
| *gloves* | 2 | 0 | 2 | 0 |

Table 8.6.6 Final reduced decision problems in "Your birthday"

**Example 8.8: Your birthday.**

The story is essentially the same as in "Barbara's birthday" after switching the roles of you and Barbara. More precisely, tomorrow it is your birthday, and Barbara wants to buy you a present. She can choose between buying a *hat,* a pair of *gloves,* or a *scarf.* She prefers buying a *hat* to buying *gloves,* which she prefers to buying a *scarf,* but at the same time she would like to surprise you. You, on the other hand, try to guess which present she bought. This situation can be represented by the psychological game in Table 8.6.5. In particular, your preferences only depend on your first-order belief, whereas Barbara's preferences only depend on her second-order belief.

**Question 8.6.1** *Draw the causality diagram, from player 1's viewpoint, that belongs to this game.*

It may be verified that this causality diagram does not induce conflicting conditions on higher-order beliefs. As such, we suspect that the choices you can rationally make under common belief in rationality are precisely those that survive the *iterated elimination of choices and states.*

To show that this is indeed true, let us first apply the procedure to the game above. Since this game is essentially equivalent to "Barbara's birthday" after switching the roles of you and Barbara, it should come as no surprise that the final reduced decision problems in the procedure are given by those in Table 8.6.6. Please verify this. We will see that under common belief in rationality, you can indeed rationally guess *hat* and *gloves.*

Like in the example "Barbara's birthday" we proceed in steps. On the basis of Barbara's final reduced decision problem in Table 8.6.6, we start by constructing a partial beliefs diagram where every surviving choice of Barbara is supported by a second-order belief that only points to surviving choices of Barbara. Such a partial beliefs diagram is given in Figure 8.6.5.

In the next step we will fill in Barbara's first-order beliefs, in a similar way as we have done for "Barbara's birthday". Consider Barbara's choice *hat,* which is optimal for the second-order belief where she believes that you believe that she will buy *gloves.* If, in addition, she believes in your rationality, then she must believe that you will guess *gloves.* Thus, Barbara's choice *hat* can be supported by the combination of a first- and second-order belief, where she believes that you will guess *gloves,* and she believes that you believe that she will buy *gloves.* Moreover, this combination will be consistent with Barbara's belief in your rationality, and Barbara's belief that you believe in Barbara's rationality.

Similarly, Barbara's choice *gloves* can be supported by the combination of a first- and second-order belief where she believes that you will guess *hat,* and she believes that you believe that she will buy

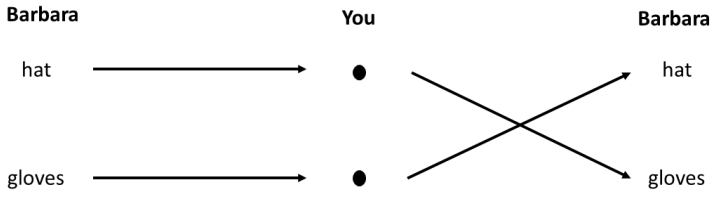



Figure 8.6.5 Partial beliefs diagram for "Your birthday"

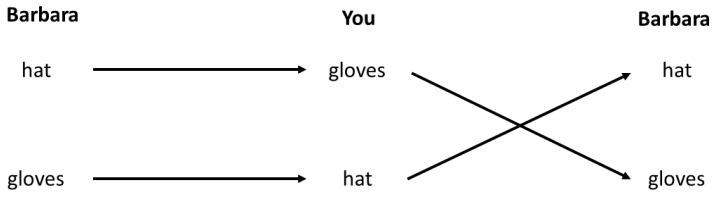



Figure 8.6.6 Full beliefs diagram for "Your birthday"

a *hat.* This leads to the full beliefs diagram in Figure 8.6.6.

From this beliefs diagram it can immediately be seen that you can rationally guess *gloves* and *hat* under common belief in rationality. We therefore conclude that the choices you can rationally make under common belief in rationality are precisely the choices that survive the *iterated elimination of choices and states.*

Such a construction of the partial and full beliefs diagram is always possible whenever player 1's preferences only depend on his first-order belief and player 2's preferences only depend on his second-order belief. As a consequence, for every such game we can always find player 1's choices that can rationally be made under common belief in rationality by performing the *iterated elimination of choices and states.*

Above we have seen that this is also true for every game where player 1's preferences only depend on his second-order belief. Finally, if player 1's and player 2's preferences only depend on their first-order belief then we are dealing with a standard game, as defined in Chapter 3, for which we know that the *iterated elimination of choices and states* yields precisely the choices that are possible under common belief in rationality. Altogether we thus arrive at the following conclusion.

**Theorem 8.6.1 (When elimination of choices and states is sufficient)** *Consider a psychological game where either*

*(i) player 1's preferences only depend on his second-order belief, or*

*(ii) player 1's and player 2's preferences only depend on their first-order belief, or*

*(iii) player 1's preferences only depend on his first-order belief and player 2's preferences only depend on his second-order belief.*

*Then, the choices that player 1 can rationally make under common belief in rationality are precisely those that survive the iterated elimination of choices and states.*

In other words, if we wish to find player 1's choices that can rationally be made under common belief in rationality, then the only instances where we need the full-fledged *iterated elimination of choices and second-order expectations* are the situations where player 1's preferences depend on his first- and second-order belief, and the situations where player 1's preferences depend only on his first-order belief and player 2's preferences depend on his first- and second-order belief. In all other situations we can safely use the easier procedure of *iterated elimination of choices and states.*

# 8.7 Proofs

## 8.7.1 Proofs of Section 8.4

Contrary to the order of the theorems in Section 8.4, we start by proving Theorem 8.4.2.

**Proof of Theorem 8.4.2.** For every round $k$ and player $i$, let $C_i^k$ be the set of choices for player $i$ that survive round $k$, and $E_i^k$ the set of second-order expectations for player $i$ that survive round $k$. For every choice $c_i$ and round $k \geq 2$ we denote by $E_i^k(c_i)$ the set of second-order expectations in $E_i^k$ for which the choice $c_i$ is optimal. Finally, for every round $k \geq 2$ we denote by $B_i^k(c_i)$ the set of first-order beliefs induced by the second-order expectations in $E_i^{k-1}(c_i)$. We define $E_i^1$ to be the set of all second-order expectations, $E_i^1(c_i)$ to be the set of second-order expectations in $E_i^1$ for which $c_i$ is optimal, and $B_i^1(c_i)$ to be the set of all first-order beliefs. Moreover, we define $C_i^0$ to be the set of all choices for player $i$. We show the following preparatory result.

*Claim.* For every player $i$ and round $k \geq 1$, the set $B_i^k(c_i)$ is non-empty for every $c_i \in C_i^{k-1}$, the sets $E_i^k$ and $C_i^k$ are non-empty, and the set $E_i^k(c_i)$ is non-empty for every $c_i \in C_i^k$.

*Proof of claim.* By induction on $k$.

**Induction start.** We start with $k = 1$. Take a choice $c_i \in C_i^0 = C_i$. Then, by definition, $B_i^1(c_i)$ is the set of all first-order beliefs, and is thus non-empty. Moreover, $E_i^1$ is the set of all second-order expectations, which is non-empty as well. To show that $C_i^1$ is non-empty, take a choice $c_i$ that is optimal for some second-order expectation in the decision problem for player $i$. Then, by Theorem 2.6.1, the choice $c_i$ is not strictly dominated, and hence $c_i \in C_i^1$. In particular, $C_i^1$ is non-empty. Finally, take some choice $c_i \in C_i^1$. Then, $c_i$ is not strictly dominated in $i$'s decision problem, which implies by Theorem 2.6.1 that $c_i$ is optimal for some second-order expectation $e_i \in E_i^1$. As such, $e_i \in E_i^1(c_i)$, which means that $E_i^1(c_i)$ is non-empty.

**Induction step.** Let $k \geq 2$, and suppose that the claim is true for $k-1$. Take some choice $c_i \in C_i^{k-1}$. Since $B_i^k(c_i)$ is the set of first-order beliefs induced by the second-order expectations in $E_i^{k-1}(c_i)$ and, by the induction assumption, $E_i^{k-1}(c_i)$ is non-empty, it follows that $B_i^k(c_i)$ is non-empty as well. The same applies for opponent $j$, and hence $B_j^k(c_j)$ is non-empty for every $c_j \in C_j^{k-1}$. As such, we can construct a second-order expectation $e_i$ that only assigns positive probability to opponent's choices in $C_j^{k-1}$, and where

$$e_i(\cdot \mid c_j) \in B_j^k(c_j)$$

for every $c_j$ to which $e_i$ assigns positive probability. By construction, such $e_i$ is in $E_i^k$, and therefore $E_i^k$ is non-empty. Next, take a choice $c_i$ that is optimal for a second-order expectation in $E_i^k$. Then, by definition, $c_i \in C_i^k$, which means that $C_i^k$ is non-empty as well. Finally, take some $c_i \in C_i^k$. Then, by definition, there is some $e_i \in E_i^k$ for which $c_i$ is optimal. As such, $e_i \in E_i^k(c_i)$. In particular, $E_i^k(c_i)$ is non-empty for all $c_i \in C_i^k$.

By induction on $k$, it follows that the claim is true for every round $k$. In particular, we conclude that $C_i^k$ is non-empty for every round $k$. Since $C_i^k \subseteq C_i^{k-1}$ for every $k \geq 1$, and there are only finitely many choices, there must for every player $i$ be a choice $c_i$ that is in $C_i^k$ for every round $k$. But then, this choice $c_i$ survives all rounds of the procedure. Thus, for every player $i$ there is at least one choice that survives all rounds. This completes the proof. ■

For the proof of Theorem 8.4.1 we need the following preparatory result, which is known in mathematics as *Cantor's intersection theorem.* Recall from Section 2.7.3 what it means for a set to be *closed* and *bounded.*

**Lemma 8.7.1 (Cantor's intersection theorem)** *Let $X$ be a finite set, and let $A_1, A_2, A_3, ...$ be an infinite sequence of nested subsets of $\mathbf{R}^X$ such that (a) $A_{k+1} \subseteq A_k$ for every $k \geq 1$, and (b) $A_k$ is non-empty, closed and bounded for every $k$. Then, there is some $v \in \mathbf{R}^X$ such that $v \in A_k$ for every $k \geq 1$.*

The proof can be found in many books on mathematics.

**Proof of Theorem 8.4.1.** For every round $k \geq 2$, recall that $E_i^k$ is the set of second-order expectations for player $i$ that survive round $k$ of the procedure. For round 1, we define $E_i^1$ to be the set of all second-order expectations. Moreover, we say that every type $t_i$ expresses 0-fold belief in rationality. In other words, 0-fold belief in rationality imposes no restrictions on types.

In order to prove this theorem we first show the following preparatory result, which states that the second-order expectations that survive round $k$ are precisely those that are consistent with up to $(k-1)$-fold belief in rationality.

*Claim.* For every $k \geq 1$ and every player $i$, we have that $e_i \in E_i^k$ if and only if there is an epistemic model, and a type $t_i$ for player $i$ in it, such that $t_i$ induces $e_i$ and $t_i$ expresses up to $(k-1)$-fold belief in rationality.

*Proof of claim.* We show the statement by induction on $k$.

**Induction start.** We start with $k = 1$. Take some $e_i \in E_i^1$. Then, there is an epistemic model, and a type $t_i$ within it, that induces the second-order expectation $e_i$. Moreover, $t_i$ satisfies 0-fold belief in rationality.

Suppose next that there is an epistemic model, and a type $t_i$ for player $i$ within it, that induces $e_i$ and expresses 0-fold belief in rationality. Then, trivially, $e_i \in E_i^1$ as $E_i^1$ contains all second-order expectations. This completes the induction start.

**Induction step.** Suppose that $k \geq 2$ and that the claim is true for $k-1$. Suppose first that $e_i \in E_i^k$. We show that there is a type $t_i$ that expresses up to $(k-1)$-fold belief in rationality and that induces the second-order expectation $e_i$. Since $e_i \in E_i^k$ we know, by definition, that

$$e_i(\cdot \mid c_j) \in B_j^k(c_j) \tag{8.7.1}$$

whenever $e_i$ assigns positive probability to the choice $c_j$. Let $C_j^+(e_i)$ be the set of choices for player $j$ that receive positive probability by $e_i$. Moreover, let $e_i(c_j)$ denote the probability that $e_i$ assigns to the opponent's choice $c_j$.

By definition, there is for every first-order belief $b_j^1 \in B_j^k(c_j)$ a second-order expectation $e_j \in E_j^{k-1}$ that induces $b_j^1$ and for which the choice $c_j$ is optimal. Hence, in view of (8.7.1), there is for every choice $c_j \in C_j^+(e_i)$ some second-order expectation $e_j[c_j] \in E_j^{k-1}$ that induces the first-order belief $e_i(\cdot \mid c_j)$ and for which the choice $c_j$ is optimal.

By the induction assumption, there is for every choice $c_j \in C_j^+(e_i)$ an epistemic model $M[c_j]$, and a type $t_j[c_j]$ in $M[c_j]$, such that $t_j[c_j]$ expresses up to $(k-2)$-fold belief in rationality, and such that $t_j[c_j]$ has the second-order expectation $e_j[c_j] \in E_j^{k-1}$. Create a large epistemic model that contains all these epistemic models $M[c_j]$ for every $c_j \in C_j^+(e_i)$. Moreover, define a new type $t_i$ within this epistemic model such that

$$b_i(t_i)(c_j, t_j) := \begin{cases} e_i(c_j), & \text{if } c_j \in C_j^+(e_i) \text{ and } t_j = t_j[c_j] \\ 0, & \text{otherwise} \end{cases} . \tag{8.7.2}$$

Then, $t_i$ only assigns positive probability to choice-type pairs $(c_j, t_j[c_j])$, where $c_j \in C_j^+(e_i)$. By construction, every such type $t_j[c_j]$ induces the second-order expectation $e_j[c_j]$ for which the choice $c_j$ is optimal. Hence, $t_i$ only assigns positive probability to choice-type pairs $(c_j, t_j[c_j])$ where the choice $c_j$ is optimal for the type $t_j[c_j]$. As such, $t_i$ expresses 1-fold belief in rationality.

Moreover, by (8.7.2), the type $t_i$ only assigns positive probability to opponent's types $t_j[c_j]$ where $c_j \in C_j^+(e_i)$. Since every such type $t_j[c_j]$ expresses up to $(k-2)$-fold belief in rationality, it follows that type $t_i$ expresses 2-fold up to $(k-1)$-fold belief in rationality. Since we already saw that $t_i$ expresses 1-fold belief in rationality, we conclude that $t_i$ expresses up to $(k-1)$-fold belief in rationality.

We next show that $t_i$ induces the second-order expectation $e_i$. Let $e_i[t_i]$ be the second-order expectation induced by the type $t_i$. By (8.7.2) we know that $e_i[t_i]$ assigns probability $e_i(c_j)$ to every choice $c_j \in C_j^+(e_i)$ and probability zero to all other choices. Hence, $e_i[t_i]$'s first-order belief coincides with the first-order belief of $e_i$.

Now, consider a choice $c_j \in C_j^+(e_i)$ and the associated type $t_j[c_j]$. Recall that type $t_j[c_j]$ induces the second-order expectation $e_j[c_j]$ which, in turn, induces the first-order belief $e_i(\cdot \mid c_j)$. Hence, the type $t_j[c_j]$ holds the first-order belief $e_i(\cdot \mid c_j)$. In view of (8.7.2) we can thus conclude that

$$(e_i[t_i])(\cdot \mid c_j) = e_i(\cdot \mid c_j) \tag{8.7.3}$$

for every $c_j \in C_j^+(e_i)$.

By (8.7.3) and the insight above that $e_i[t_i]$ has the same first-order belief as $e_i$, we conclude that $e_i[t_i] = e_i$. That is, type $t_i$ induces the second-order expectation $e_i$. Altogether, we see that there is a type $t_i$ that expresses up to $(k-1)$-fold belief in rationality and that induces the second-order expectation $e_i$. Hence, for every $e_i \in E_i^k$ there is a type $t_i$ that expresses up to $(k-1)$-fold belief in rationality and that induces the second-order expectation $e_i$.

Conversely, suppose that $e_i$ is a second-order expectation for which there is a type $t_i$ that expresses up to $(k-1)$-fold belief in rationality and that induces the second-order expectation $e_i$. We will show that $e_i \in E_i^k$.

Let $(c_j, t_j)$ be an opponent's choice-type pair with $b_i(t_i)(c_j, t_j) > 0$. Since $t_i$ expresses up to $(k-1)$-fold belief in rationality, we conclude that $c_j$ is optimal for $t_j$, and that $t_j$ expresses up to $(k-2)$-fold belief in rationality. Let $e_j[t_j]$ be the second-order expectation induced by $t_j$. As $t_j$ expresses up to $(k-2)$-fold belief in rationality, we know by the induction assumption that $e_j[t_j] \in E_j^{k-1}$. Hence, we conclude that

$$b_i(t_i)(c_j, t_j) > 0 \text{ only if } c_j \text{ is optimal for } t_j \text{ and } e_j[t_j] \in E_j^{k-1}.$$

For a given choice $c_j$, let $E_j^{k-1}(c_j)$ be the set of second-order expectations in $E_j^{k-1}$ for $c_j$ is optimal. Together with our insight above, we thus see that

$$b_i(t_i)(c_j, t_j) > 0 \text{ only if } e_j[t_j] \in E_j^{k-1}(c_j).$$

Since, by definition, $B_j^k(c_j)$ are the first-order beliefs induced by the second-order expectations in $E_j^{k-1}(c_j)$, we conclude that

$$b_i(t_i)(c_j, t_j) > 0 \text{ only if } b_j^1[t_j] \in B_j^k(c_j), \tag{8.7.4}$$

where $b_j^1[t_j]$ is the first-order belief induced by the type $t_j$.

On the basis of (8.7.4) we will now show that

$$e_i(\cdot \mid c_j) \in B_j^k(c_j) \tag{8.7.5}$$

for every $c_j \in C_j^+(e_i)$. Recall that $e_i$ is the second-order expectation induced by the type $t_i$. Take some $c_j \in C_j^+(e_i)$. Then, we have that

$$e_i(c_i \mid c_j) = \frac{\sum_{t_j : b_i(t_i)(c_j,t_j)>0} b_i(t_i)(c_j,t_j) \cdot b_j^1[t_j](c_i)}{b_i^1[t_i](c_j)}$$

for all $c_i \in C_i$. This implies that

$$e_i(\cdot \mid c_j) = \sum_{t_j : b_i(t_i)(c_j,t_j)>0} \lambda[t_j] \cdot b_j^1[t_j], \tag{8.7.6}$$

where

$$\lambda[t_j] := \frac{b_i(t_i)(c_j,t_j)}{b_i^1[t_i](c_j)} \text{ for every } t_j \text{ with } b_i(t_i)(c_j,t_j) > 0. \tag{8.7.7}$$

From (8.7.7) we see that $\lambda[t_j] \geq 0$, and

$$\sum_{t_j : b_i(t_i)(c_j,t_j)>0} \lambda[t_j] = 1.$$

Together with (8.7.6) we thus conclude that $e_i(\cdot \mid c_j)$ is a convex combination of the first-order beliefs $b_j^1[t_j]$, where $t_j$ is such that $b_i(t_i)(c_j,t_j) > 0$. By (8.7.4) we know that $b_j^1[t_j] \in B_j^k(c_j)$ for every type $t_j$ with $b_i(t_i)(c_j,t_j) > 0$. Together with our insight above, we see that

$$e_i(\cdot \mid c_j) \text{ is a convex combination of first-order beliefs } b_j^1[t_j] \in B_j^k(c_j). \tag{8.7.8}$$

We now show that $E_j^{k-1}(c_j)$ is a convex set. Take some $e_j, \hat{e}_j \in E_j^{k-1}(c_j)$ and some $\lambda \in [0,1]$. We show that $(1-\lambda) \cdot e_j + \lambda \cdot \hat{e}_j \in E_j^{k-1}(c_j)$. By definition of $E_j^{k-1}(c_j)$, the choice $c_j$ is optimal for $e_j$ and $\hat{e}_j$. But then, $c_j$ is optimal for $(1-\lambda) \cdot e_j + \lambda \cdot \hat{e}_j$. Indeed, for every $c_j' \in C_j$ we have that

$$\begin{aligned} u_j(c_j, (1-\lambda) \cdot e_j + \lambda \cdot \hat{e}_j) &= (1-\lambda) \cdot u_j(c_j, e_j) + \lambda \cdot u_j(c_j, \hat{e}_j) \\ &\geq (1-\lambda) \cdot u_j(c_j', e_j) + \lambda \cdot u_j(c_j', \hat{e}_j) \\ &= u_j(c_j', (1-\lambda) \cdot e_j + \lambda \cdot \hat{e}_j), \end{aligned}$$

where the inequality follows the fact that $c_j$ is optimal for $e_j$ and $\hat{e}_j$. We therefore see that $c_j$ is optimal for $(1-\lambda) \cdot e_j + \lambda \cdot \hat{e}_j$. Hence, $(1-\lambda) \cdot e_j + \lambda \cdot \hat{e}_j \in E_j^{k-1}(c_j)$. As such, we conclude that the set $E_j^{k-1}(c_j)$ is convex.

By definition, $B_j^k(c_j)$ is the set of first-order beliefs implied by the second-order expectations in $E_j^{k-1}(c_j)$. Since the set $E_j^{k-1}(c_j)$ is convex, it follows that the set $B_j^k(c_j)$ is convex as well.

But then, (8.7.8) implies that $e_i(\cdot \mid c_j) \in B_j^k(c_j)$. Since this holds for every $c_j \in C_j^+(e_i)$, we conclude that (8.7.5) holds. Hence, by definition, $e_i \in E_i^k$.

We have thus shown that every second-order expectation $e_i$ for which there is a type $t_i$ that expresses up to $(k-1)$-fold belief in rationality and that induces the second-order expectation $e_i$, must be in $E_i^k$. As the converse has also been shown above, we conclude that the statement of the claim holds for $k$. By induction on $k$, the statement in the claim holds for every $k$. This completes the proof of the claim.

We are now able to prove parts (a) and (b) of Theorem 8.4.1.

**(a)** Suppose that player $i$ can rationally make the choice $c_i$ while expressing up to $k$-fold belief in rationality. Then, there is a type $t_i$ that expresses up to $k$-fold belief in rationality and for which the choice $c_i$ is optimal. Suppose that $t_i$ induces the second-order expectation $e_i$. Then, we now by the claim that $e_i \in E_i^{k+1}$. Moreover, $c_i$ is optimal for $e_i$. As such, $c_i$ survives the first $k+1$ rounds of the procedure.

Suppose next that choice $c_i$ survives the first $k+1$ rounds of the procedure. Then, by definition, there is a second-order expectation $e_i \in E_i^{k+1}$ for which $c_i$ is optimal. By the claim we know that there is a type $t_i$ that expresses up to $k$-fold belief in rationality and that induces the second-order expectation $e_i$. Since $c_i$ is optimal for $e_i$, we know that $c_i$ is also optimal for $t_i$, and hence $c_i$ can rationally be made while expressing up to $k$-fold belief in rationality.

**(b)** Suppose that choice $c_i$ can rationally be made while expressing common belief in rationality. Then, by (a) we know that $c_i$ survives the first $k+1$ rounds, for every $k \geq 1$, and hence $c_i$ survives all rounds of the procedure.

We will now construct an epistemic model where, for every choice $c_i$ that survives all rounds of the procedure, there is a type that expresses common belief in rationality and for which the choice $c_i$ is optimal.

For every choice $c_i$ and round $k$, let $E_i^k(c_i)$ denote the set of second-order expectations in $E_i^k$ for which the choice $c_i$ is optimal. To construct the epistemic model we look, for both players $i$, at the set of choices $C_i^*$ that survive all rounds of the procedure, at the set

$$E_i^* := \{e_i \in E_i \mid e_i \in E_i^k \text{ for all } k \geq 1\},$$

and for every choice $c_i \in C_i^*$ at the set

$$E_i^*(c_i) := \{e_i \in E_i \mid e_i \in E_i^k(c_i) \text{ for all } k \geq 1\}.$$

Hence, $E_i^*$ contains all the second-order expectations that survive all the rounds of the procedure, whereas $E_i^*(c_i)$ contains all the second-order expectations in $E_i^*$ that support the choice $c_i$.

We will now show that the set $E_i^*(c_i)$ is non-empty for every choice $c_i \in C_i^*$. Take some $c_i \in C_i^*$ that survives all rounds. Then, for every round $k$ there is second-order expectation $e_i^k \in E_i^k(c_i)$ for which the choice $c_i$ is optimal. We will show, by means of Lemma 8.7.1, that there is some second-order expectation $e_i$ which is in $E_i^k(c_i)$ for all $k$, and therefore is in $E_i^*(c_i)$.

By construction, $E_i^k(c_i)$ consists of probability distributions on $C_j \times C_i$, and hence $E_i^k(c_i)$ is a subset of $\mathbf{R}^{C_j \times C_i}$. Since for every $k$ there is some $e_i^k$ in $E_i^k(c_i)$, we conclude that every set $E_i^k(c_i)$ is non-empty. Moreover, since every $e_i \in E_i^k(c_i)$ corresponds to a vector in $\mathbf{R}^{C_j \times C_i}$ where all coordinates are between 0 and 1, we conclude that every set $E_i^k(c_i)$ is bounded.

We next show that $E_i^k(c_i)$ is closed, by induction on $k$.

For $k = 1$ we have that $E_i^k(c_i) = E_i = \Delta(C_j \times C_i)$, which is a closed set.

Now, take some $k \geq 2$, and assume that $E_i^{k-1}(c_i)$ is closed. We show that $E_i^k(c_i)$ is closed, by showing that $\mathbf{R}^{C_j \times C_i} \backslash E_i^k(c_i)$ is open. Take some $v_i \in \mathbf{R}^{C_j \times C_i} \backslash E_i^k(c_i)$. Then, either (i) $v_i$ is not a probability distribution on $C_j \times C_i$, or (ii) $v_i$ is a probability distribution but $c_i$ is not optimal for $v_i$, or (iii) $v_i$ is a probability distribution, $c_i$ is optimal for $v_i$, but $v_i \notin E_i^k(c_i)$. In cases (i) and (ii) we can find an open ball $B_r(v_i)$ around $v_i$ such that $B_r(v_i) \subseteq \mathbf{R}^{C_j \times C_i} \backslash E_i^k(c_i)$.

In case (iii) we know that there is some $c_j$ with

$$v_i(c_j) > 0 \text{ such that } v_i(\cdot \mid c_j) \notin B_j^k(c_j). \tag{8.7.9}$$

By the induction assumption we know that $E_j^{k-1}(c_j)$ is closed. As $B_j^k(c_j)$ contains the first-order beliefs induced by the second-order expectations in $E_j^{k-1}(c_j)$, it follows that $B_j^k(c_j)$ is closed as well. By (8.7.9) we conclude that there is an open ball $B_r(v_i)$ around $v_i$ such that (8.7.9) holds for every $w_i \in B_r(v_i)$. Hence, there is an open ball $B_r(v_i)$ around $v_i$ such that $B_r(v_i) \subseteq \mathbf{R}^{C_j \times C_i} \backslash E_i^k(c_i)$.

In view of the cases (i), (ii) and (iii), it follows that the set $\mathbf{R}^{C_j \times C_i} \backslash E_i^k(c_i)$ is open, which means that $E_i^k(c_i)$ is closed.

Finally, by construction, we have that $E_i^{k+1}(c_i) \subseteq E_i^k(c_i)$ for every $k \geq 1$. We may therefore conclude, by Lemma 8.7.1, that there is some second-order expectation $e_i$ such that $e_i \in E_i^k(c_i)$ for all $k$. As such, $e_i \in E_i^*(c_i)$, and thus $E_i^*(c_i)$ is non-empty for every choice $c_i \in C_i^*$. Since we know, from Theorem 8.4.2, that $C_i^*$ is non-empty, this implies that $E_i^*$ is non-empty as well.

For every choice $c_i$, let $B_i^*(c_i)$ be the set of first-order beliefs for player $i$ that are induced by some second-order expectation in $E_i^*(c_i)$. By construction of the procedure, we then have that $e_i \in E_i^*$ precisely when

$$e_i(\cdot \mid c_j) \in B_j^*(c_j) \tag{8.7.10}$$

for every $c_j$ to which $e_i$ assigns positive probability. On the basis of this property we can now construct second-order expectations as follows.

We first select, for both players $i$ and every choice $c_i^1 \in C_i^*$, a second-order expectation $e_i[c_i^1] \in E_i^*(c_i^1)$ for which $c_i^1$ is optimal.

Now, take some $c_i^1 \in C_i^*$ and some $c_j^2$ to which $e_i[c_i^1]$ assigns positive probability. Then, we have by (8.7.10) that $e_i[c_i^1](\cdot \mid c_j^2) \in B_j^*(c_j^2)$. This means that there must be some second-order expectation $e_j[c_i^1, c_j^2] \in E_j^*(c_j^2)$ that has the first-order belief $e_i[c_i^1](\cdot \mid c_j^2)$ for which $c_j^2$ is optimal. In particular, it follows that $c_j^2 \in C_j^*$. Hence, for every $c_i^1 \in C_i^*$ and every $c_j^2 \in C_j^*$ to which $e_i[c_i^1]$ assigns positive probability, there is some second-order expectation $e_j[c_i^1, c_j^2] \in E_j^*(c_j^2)$ for which $c_j^2$ is optimal and that has the first-order belief $e_i[c_i^1](\cdot \mid c_j^2)$.

Next, take some $c_i^1 \in C_i^*$, some $c_j^2 \in C_j^*$ to which $e_i[c_i^1]$ assigns positive probability, and some $c_i^3$ to which $e_j[c_i^1, c_j^2]$ assigns positive probability. As $e_j[c_i^1, c_j^2] \in E_j^*$, it follows by (8.7.10) that $e_j[c_i^1, c_j^2](\cdot \mid c_i^3) \in B_i^*(c_i^3)$. This means that there must be some second-order expectation $e_i[c_i^1, c_j^2, c_i^3] \in E_i^*(c_i^3)$ for which $c_i^3$ is optimal and that has the first-order belief $e_j[c_i^1, c_j^2](\cdot \mid c_i^3)$. In particular, it follows that $c_i^3 \in C_i^*$. Hence, for every $c_i^1 \in C_i^*$, every $c_j^2 \in C_j^*$ to which $e_i[c_i^1]$ assigns positive probability, and every $c_i^3 \in C_i^*$ to which $e_j[c_i^1, c_j^2]$ assigns positive probability, there is some second-order expectation $e_i[c_i^1, c_j^2, c_i^3] \in E_i^*(c_i^3)$ that has the first-order belief $e_j[c_i^1, c_j^2](\cdot \mid c_i^3)$ and for which $c_i^3$ is optimal.

By continuing in this fashion, we can construct for every odd $m \geq 1$, every $c_i^1 \in C_i^*, c_j^2 \in C_j^*$, ... , $c_i^m \in C_i^*$ a second-order expectation $e_i[c_i^1, c_j^2, ..., c_i^m]$, and for every even $m \geq 2$, every $c_j^1 \in C_j^*$, $c_i^2 \in C_j^*$, ... , $c_i^m \in C_i^*$ a second-order expectation $e_i[c_j^1, c_i^2, ..., c_i^m]$ such that the following conditions hold: For every odd $m \geq 1$,

$$c_i^m \text{ is optimal for } e_i[c_i^1, c_j^2, ..., c_i^m] \tag{8.7.11}$$

and

$$e_i[c_i^1, c_j^2, ..., c_i^m] \text{ has the first-order belief } e_j[c_i^1, c_j^2, ..., c_j^{m-1}](\cdot \mid c_i^m) \tag{8.7.12}$$

whenever $e_i[c_i^1]$ assigns a positive probability to $c_j^2$, $e_j[c_i^1, c_j^2]$ assigns a positive probability to $c_i^3$, ... , $e_j[c_i^1, c_j^2, ..., c_j^{m-1}]$ assigns a positive probability to $c_i^m$. Moreover, for every even $m \geq 2$,

$$c_i^m \text{ is optimal for } e_i[c_j^1, c_i^2, ..., c_i^m] \tag{8.7.13}$$

and

$$e_i[c_j^1, c_i^2, ..., c_i^m] \text{ has the first-order belief } e_j[c_j^1, c_i^2, ..., c_j^{m-1}](\cdot \mid c_i^m) \tag{8.7.14}$$

whenever $e_j[c_j^1]$ assigns a positive probability to $c_i^2$, $e_i[c_j^1, c_i^2]$ assigns a positive probability to $c_j^3$, ... , $e_j[c_j^1, c_i^2, ..., c_j^{m-1}]$ assigns a positive probability to $c_i^m$.

On the basis of the properties (8.7.11), (8.7.12), (8.7.13) and (8.7.14), we can now construct an epistemic model as follows. For both players $i$, let $D_i$ be the collection of sequences $[c_i^1, c_j^2, ..., c_i^m]$ above for odd $m$ where $c_i^1 \in C_i^*$, $c_j^2 \in C_j^*, ..., c_i^m \in C_i^*$, such that $e_i[c_i^1]$ assigns a positive probability to $c_j^2$, $e_j[c_i^1, c_j^2]$ assigns a positive probability to $c_i^3$, ... , $e_j[c_i^1, c_j^2, ..., c_j^{m-1}]$ assigns a positive probability to $c_i^m$ (if $m \geq 3$), together with the collection of sequences $[c_j^1, c_i^2, ..., c_i^m]$ above for even $m$ where $c_j^1 \in C_j^*$, $c_i^2 \in C_i^*, ..., c_i^m \in C_i^*$, such that $e_j[c_j^1]$ assigns a positive probability to $c_i^2$, $e_i[c_j^1, c_i^2]$ assigns a positive probability to $c_j^3$, ... , $e_j[c_i^1, c_j^2, ..., c_j^{m-1}]$ assigns a positive probability to $c_i^m$.

For every odd $m \geq 1$, both players $i$, and every $[c_i^1, c_j^2, ..., c_i^m] \in D_i$, we define a type $t_i[c_i^1, c_j^2, ..., c_i^m]$, and for every even $m \geq 2$, both players $i$, and every $[c_j^1, c_i^2, ..., c_i^m] \in D_i$, we define a type $t_i[c_j^1, c_i^2, ..., c_i^m]$ with the following beliefs: For every odd $m \geq 1$,

$$b_i(t_i[c_i^1, c_j^2, ..., c_i^m])(c_j, t_j) := \begin{cases} e_i[c_i^1, c_j^2, ..., c_i^m](c_j), & \text{if } [c_i^1, c_j^2, ..., c_i^m, c_j] \in D_j \\ & \text{and } t_j = t_j[c_i^1, c_j^2, ..., c_i^m, c_j] \\ 0, & \text{otherwise} \end{cases} \tag{8.7.15}$$

for every opponent's choice-type pair $(c_j, t_j)$. Here, $e_i[c_i^1, c_j^2, ..., c_i^m](c_j)$ denotes the probability that the second-order expectation $e_i[c_i^1, c_j^2, ..., c_i^m]$ assigns to the choice $c_j$. Moreover, for every even $m \geq 2$,

$$b_i(t_i[c_j^1, c_i^2, ..., c_i^m])(c_j, t_j) := \begin{cases} e_i[c_j^1, c_i^2, ..., c_i^m](c_j), & \text{if } [c_j^1, c_i^2, ..., c_i^m, c_j] \in D_j \\ & \text{and } t_j = t_j[c_j^1, c_i^2, ..., c_i^m, c_j] \\ 0, & \text{otherwise} \end{cases} \tag{8.7.16}$$

for every opponent's choice-type pair $(c_j, t_j)$. This completes the construction of the epistemic model. Note that the epistemic model contains infinitely many types, since the sets $D_i$ and $D_j$ are infinite.

We first show that every type $t_i[c_i^1, c_j^2, ..., c_i^m]$ induces the second-order expectation $e_i[c_i^1, c_j^2, ..., c_i^m]$. Let $e_i[t_i[c_i^1, c_j^2, ..., c_i^m]]$ be the second-order expectation induced by type $t_i$. By (8.7.15) we see that $t_i[c_i^1, c_j^2, ..., c_i^m]$ has the same first-order belief as $e_i[c_i^1, c_j^2, ..., c_i^m]$, and hence $e_i[t_i[c_i^1, c_j^2, ..., c_i^m]]$ has the same first-order belief as $e_i[c_i^1, c_j^2, ..., c_i^m]$.

Now, take some choice $c_j$ to which $t_i[c_i^1, c_j^2, ..., c_i^m]$ assigns positive probability. Then, by (8.7.15), $e_i[c_i^1, c_j^2, ..., c_i^m]$ assigns positive probability to $c_j$. Hence, by definition of $D_j$, we have that $[c_i^1, c_j^2, ..., c_i^m, c_j] \in D_j$. Moreover, by (8.7.15), there is only one type $t_j$ such that $t_i[c_i^1, c_j^2, ..., c_i^m]$ assigns positive probability to $(c_j, t_j)$, and this is the type $t_j[c_i^1, c_j^2, ..., c_i^m, c_j]$. By (8.7.16) we know that $t_j[c_i^1, c_j^2, ..., c_i^m, c_j]$ holds the same first-order belief as $e_j[c_i^1, c_j^2, ..., c_i^m, c_j]$. We thus conclude that

$$e_i[t_i[c_i^1, c_j^2, ..., c_i^m]](\cdot \mid c_j) \text{ is the first-order belief of } e_j[c_i^1, c_j^2, ..., c_i^m, c_j] \tag{8.7.17}$$

for every $c_j$ to which $e_i[t_i[c_i^1, c_j^2, ..., c_i^m]]$ assigns positive probability.

Moreover, we know by (8.7.14) that

$$e_j[c_i^1, c_j^2, ..., c_i^m, c_j] \text{ has the first-order belief } e_i[c_i^1, c_j^2, ..., c_i^m](\cdot \mid c_j) \tag{8.7.18}$$

for every $c_j$ to which $e_i[t_i[c_i^1, c_j^2, ..., c_i^m]]$ assigns positive probability. By combining (8.7.17) and (8.7.18) we conclude that

$$e_i[t_i[c_i^1, c_j^2, ..., c_i^m]](\cdot \mid c_j) = e_i[c_i^1, c_j^2, ..., c_i^m](\cdot \mid c_j) \tag{8.7.19}$$

for every $c_j$ to which $e_i[t_i[c_i^1, c_j^2, ..., c_i^m]]$ assigns positive probability. Since we already saw that $e_i[t_i[c_i^1, c_j^2, ..., c_i^m]]$ and $e_i[c_i^1, c_j^2, ..., c_i^m]$ have the same first-order belief, it follows by (8.7.19) that $e_i[t_i[c_i^1, c_j^2, ..., c_i^m]] = e_i[c_i^1, c_j^2, ..., c_i^m]$, which was to show. Hence, every type $t_i[c_i^1, c_j^2, ..., c_i^m]$ induces the second-order expectation $e_i[c_i^1, c_j^2, ..., c_i^m]$ for every odd $m$. In the same way, it can be shown that every type $t_i[c_j^1, c_i^2, ..., c_i^m]$ induces the second-order expectation $e_i[c_j^1, c_i^2, ..., c_i^m]$ for every even $m$.

Next, we prove that for every odd $m$, and every type $t_i[c_i^1, c_j^2, ..., c_i^m]$, the choice $c_i^m$ is optimal for the type $t_i[c_i^1, c_j^2, ..., c_i^m]$. We have seen above that type $t_i[c_i^1, c_j^2, ..., c_i^m]$ induces the second-order expectation $e_i[c_i^1, c_j^2, ..., c_i^m]$. Since, by (8.7.11), the choice $c_i^m$ is optimal for $e_i[c_i^1, c_j^2, ..., c_i^m]$, it follows that $c_i^m$ is indeed optimal for the type $t_i[c_i^1, c_j^2, ..., c_i^m]$. In the same way, it can be shown that for every even $m$, and every type $t_i[c_j^1, c_i^2, ..., c_i^m]$, the choice $c_i^m$ is optimal for the type $t_i[c_j^1, c_i^2, ..., c_i^m]$.

With this insight at hand, we can now show that every type in the epistemic model believes in the opponent's rationality. Consider, for an odd $m$, the type $t_i[c_i^1, c_j^2, ..., c_i^m]$. Suppose that $t_i[c_i^1, c_j^2, ..., c_i^m]$ assigns positive probability to an opponent's choice-type pair $(c_j, t_j)$. Then, it follows from (8.7.15) that $[c_i^1, c_j^2, ..., c_i^m, c_j] \in D_j$ and $t_j = t_j[c_i^1, c_j^2, ..., c_i^m, c_j]$. Since we have seen above that $c_j$ is optimal for the type $t_j[c_i^1, c_j^2, ..., c_i^m, c_j]$, we conclude that type $t_i[c_i^1, c_j^2, ..., c_i^m]$ believes in the opponent's rationality. In the same fashion it can be shown that, for every even $m$, every type $t_i[c_j^1, c_i^2, ..., c_i^m]$ believes in the opponent's rationality as well.

As every type in the epistemic model believes in the opponent's rationality, it follows by the same argument as in the proof of Theorem 3.3.1 that every type in the epistemic model expresses *common* belief in rationality.

Consider now a choice $c_i$ that survives all rounds of the procedure. Then, by definition, $c_i \in C_i^*$. As the type $t_i[c_i]$ expresses common belief in rationality and $c_i$ is optimal for the type $t_i[c_i]$, we conclude that $c_i$ can rationally be chosen under common belief in rationality. We thus see that a choice $c_i$ survives all rounds of the procedure, if and only if, $c_i$ can rationally be chosen under common belief in rationality. This completes the proof. ■

**Proof of Theorem 8.4.3.** Consider the epistemic model we constructed in the proof of Theorem 8.4.1, part (b). In that epistemic model, every type expresses common belief in rationality. ■

### 8.7.2 Proof of Section 8.6

**Proof of Theorem 8.6.1. (a)** We first show that, in every psychological game, every choice that can rationally be made under common belief in rationality must survive the *iterated elimination of choices and states.*

We use the following notation: For every round $k \geq 1$, let $C_i^k$ be the set of choices for player $i$ that survive round $k$ of the procedure. Set $C_i^0 := C_i$. Then, for every round $k$, the set $C_i^k$ contains precisely those choices that are not strictly dominated in the reduced decision problem $(C_i^{k-1}, C_j^{k-1} \times C_i^{k-1})$ where the set of choices is $C_i^{k-1}$ and the set of states only contains those pairs $(c_j, c_i)$ where $c_j \in C_j^{k-1}$ and $c_i \in C_i^{k-1}$. We show the following result.

*Claim.* Take an epistemic model $M$ and a type $t_i$ for player $i$ within it that expresses common belief in rationality. Then, for every $k \geq 0$, every choice that is optimal for $t_i$ is in $C_i^k$.

*Proof of claim.* By induction on $k$. For $k = 0$ the statement is trivially true as $C_i^0 = C_i$.

Now, take some $k \geq 1$ and assume that the statement is true for $k-1$. Take an epistemic model $M$ and a type $t_i$ within it that expresses common belief in rationality. Then, $t_i$ only assigns positive probability to opponent's choice-type pairs $(c_j, t_j)$ where $t_j$ expresses common belief in rationality and

$c_j$ is optimal for $t_j$. By the induction assumption we then know that every such choice $c_j$ is in $C_j^{k-1}$. Thus, $t_i$'s first-order belief only assigns positive probability to opponent's choices that are in $C_j^{k-1}$.

Moreover, $t_i$ only assigns positive probability to opponent's types $t_j$ which only assign positive probability to $i$'s choice-type pairs $(c_i', t_i')$ where $t_i'$ expresses common belief in rationality and $c_i'$ is optimal for $t_i'$. By the induction assumption we know that every such choice $c_i'$ is in $C_i^{k-1}$. Thus, $t_i$'s second-order belief only assigns positive probability to $i$'s choices that are in $C_i^{k-1}$.

By combining the two insights above we conclude that the second-order expectation $e_i[t_i]$ induced by $t_i$ only assigns positive probability to pairs $(c_j, c_i) \in C_j^{k-1} \times C_i^{k-1}$. Take a choice $c_i$ that is optimal for the type $t_i$. Then, $c_i$ is optimal for the second-order expectation $e_i[t_i] \in \Delta(C_j^{k-1} \times C_i^{k-1})$. By Theorem 2.6.1 it then follows that $c_i$ is not strictly dominated in the reduced decision problem $(C_i^{k-1}, C_j^{k-1} \times C_i^{k-1})$. But then, by definition of the procedure, $c_i \in C_i^k$.

By induction on $k$, the proof of the claim follows.

By the claim, we conclude that every choice that is optimal for a type that expresses common belief in rationality must survive all the rounds of *iterated elimination of choices and states.* This completes the proof of part (a).

**(b)** We now show that for every class of psychological games listed in the statement of the theorem, every choice that survives the *iterated elimination of choices and states* can rationally be made under common belief in rationality. We distinguish 5 cases: (1) player 1's and player 2's preferences only depend on the first-order belief, (2) player 1's preferences only depend on his first-order belief and player 2's preferences only depend on his second-order belief, (3) player 1's preferences only depend on his second-order belief and player 2's preferences only depend on his first-order belief, (4) player 1's and player 2's preferences only depend on the second-order belief, and (5) player 1's preferences only depend on his second-order belief and player 2's preferences depend on his first- and second-order belief.

**Case 1.** Suppose that player 1's and player 2's preferences only depend on the first-order belief. Then we are dealing with a standard game as defined in Chapter 3. Moreover, for this class of games the *iterated elimination of choices and states* coincides with the *iterated elimination of strictly dominated choices* from Chapter 3. By Theorem 3.4.1 we know that for this class of games, the choices that can rationally be made under common belief in rationality are precisely the choices that survive the *iterated elimination of strictly dominated choices,* and hence survive the *iterated elimination of choices and states.*

For Cases 2–5 we adopt the following notation: By $C_1^*$ and $C_2^*$ we denote the sets of choices for player 1 and 2, respectively, that survive the *iterated elimination of choices and states.*

**Case 2.** Suppose that player 1's preferences only depend on his first-order belief and player 2's preferences only depend on his second-order belief. Then, every choice $c_1 \in C_1^*$ is optimal for a first-order belief $b_1^{c_1} \in \Delta(C_2^*)$ and every choice $c_2 \in C_2^*$ is optimal for a second-order belief $b_2^{c_2} \in \Delta(C_2^*)$. For every second-order belief $b_2^{c_2} \in \Delta(C_2^*)$, let $c_1[c_2]$ be a choice for player 1 that is optimal if player 1 holds the belief $b_2^{c_2}$ about player 2's choice.

We construct an epistemic model with sets of types

$$T_1 = \{t_1^{c_1} \mid c_1 \in C_1^*\} \cup \{t_1^{c_2} \mid c_2 \in C_2^*\} \text{ and } T_2 = \{t_2^{c_2} \mid c_2 \in C_2^*\}.$$

The beliefs of the types are as follows:

$$b_1(t_1^{c_1})(c_2, t_2) := \begin{cases} b_1^{c_1}(c_2), & \text{if } c_2 \in C_2^* \text{ and } t_2 = t_2^{c_2} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.20}$$

for every $c_1 \in C_1^*$ and every $(c_2, t_2) \in C_2 \times T_2$,

$$b_1(t_1^{c_2})(c_2', t_2) := \begin{cases} b_2^{c_2}(c_2'), & \text{if } c_2' \in C_2^* \text{ and } t_2 = t_2^{c_2'} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.21}$$

for every $c_2 \in C_2^*$ and every $(c_2', t_2) \in C_2 \times T_2$, and

$$b_2(t_2^{c_2})(c_1, t_1) := \begin{cases} 1, & \text{if } c_1 = c_1[c_2] \text{ and } t_1 = t_1^{c_2} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.22}$$

for every $c_2 \in C_2^*$ and every $(c_1, t_1) \in C_1 \times T_1$.

By (8.7.20) we see that every type $t_1^{c_1}$ has the first-order belief $b_1^{c_1}$ for which the choice $c_1$ is optimal. From (8.7.21) we conclude that every type $t_1^{c_2}$ has the first-order belief $b_2^{c_2}$ for which the choice $c_1[c_2]$ is optimal. Moreover, by combining (8.7.21) and (8.7.22) we see that every type $t_2^{c_2}$ has the second-order belief $b_2^{c_2}$ for which the choice $c_2$ is optimal. These insights, together with (8.7.20), (8.7.21) and (8.7.22), imply that every type in the epistemic model constructed above believes in the opponent's rationality. Therefore, every type in the model expresses common belief in rationality.

Now, take an arbitrary choice $c_1 \in C_1^*$. Then, $c_1$ is optimal for the type $t_1^{c_1}$ that expresses common belief in rationality. That is, $c_1$ can rationally be chosen under common belief in rationality.

**Case 3.** Suppose that player 1's preferences only depend on his second-order belief and player 2's preferences only depend on his first-order belief. Then, every choice $c_1 \in C_1^*$ is optimal for a second-order belief $b_1^{c_1} \in \Delta(C_1^*)$ and every choice $c_2 \in C_2^*$ is optimal for a first-order belief $b_2^{c_2} \in \Delta(C_1^*)$. For every second-order belief $b_1^{c_1} \in \Delta(C_1^*)$, let $c_2[c_1]$ be a choice for player 2 that is optimal if player 2 holds the belief $b_1^{c_1}$ about player 1's choice.

We construct an epistemic model with sets of types

$$T_1 = \{t_1^{c_1} \mid c_1 \in C_1^*\} \text{ and } T_2 = \{t_2^{c_1} \mid c_1 \in C_1^*\}.$$

The beliefs of the types are as follows:

$$b_1(t_1^{c_1})(c_2, t_2) := \begin{cases} 1, & \text{if } c_2 = c_2[c_1] \text{ and } t_2 = t_2^{c_1} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.23}$$

for every $c_1 \in C_1^*$ and every $(c_2, t_2) \in C_2 \times T_2$, and

$$b_2(t_2^{c_1})(c_1', t_1) := \begin{cases} b_1^{c_1}(c_1'), & \text{if } c_1' \in C_1^* \text{ and } t_1 = t_1^{c_1'} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.24}$$

for every $c_1 \in C_1^*$ and every $(c_1', t_1) \in C_1 \times T_1$.

By combining (8.7.23) and (8.7.24) we see that every type $t_1^{c_1}$ has the second-order belief $b_1^{c_1}$ for which the choice $c_1$ is optimal. From (8.7.24) we conclude that every type $t_2^{c_1}$ has the first-order belief $b_1^{c_1}$ for which the choice $c_2[c_1]$ is optimal. These insights, together with (8.7.23) and (8.7.24), imply that every type in the epistemic model constructed above believes in the opponent's rationality. Therefore, every type in the model expresses common belief in rationality.

Now, take an arbitrary choice $c_1 \in C_1^*$. Then, $c_1$ is optimal for the type $t_1^{c_1}$ that expresses common belief in rationality. That is, $c_1$ can rationally be chosen under common belief in rationality.

**Case 4.** Suppose that player 1's and player 2's preferences only depend on the second-order belief. Then, every choice $c_1 \in C_1^*$ is optimal for a second-order belief $b_1^{c_1} \in \Delta(C_1^*)$. Consider a choice $c_2^1 \in C_2^*$

for player 2. For every $k \geq 2$, let $c_2^k$ be a choice that is optimal for player 2 if he believes that player 1 assigns probability 1 to player 2 choosing $c_2^{k-1}$. Since there are finitely many choices, there must be some $k \geq 1$ and $m \geq 1$ such that $c_2^{k+m} = c_2^k$.

We construct an epistemic model with sets of types

$$T_1 = \{t_1^{c_1,k+l} \mid c_1 \in C_1^*,\ l \in \{0,...,m-1\}\} \text{ and } T_2 = \{t_2^{c_1,k+l} \mid c_1 \in C_1^*,\ l \in \{0,...,m-1\}\}.$$

The beliefs of the types are as follows:

$$b_1(t_1^{c_1,k+l})(c_2,t_2) := \begin{cases} 1, & \text{if } c_2 = c_2^{k+l} \text{ and } t_2 = t_2^{c_1,k+l} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.25}$$

for every $c_1 \in C_1^*$, $l \in \{0,...,m-1\}$ and every $(c_2,t_2) \in C_2 \times T_2$, and

$$b_2(t_2^{c_1,k+l})(c_1',t_1) := \begin{cases} b_1^{c_1}(c_1'), & \text{if } c_1' \in C_1^* \text{ and } t_1 = t_1^{c_1',k+l-1} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.26}$$

for every $c_1 \in C_1^*$, $l \in \{0,...,m-1\}$ and every $(c_1',t_1) \in C_1 \times T_1$. Here, we use the convention that $t_1^{c_1',k-1} = t_1^{c_1',k+m-1}$.

By combining (8.7.25) and (8.7.26) we see that every type $t_1^{c_1,k+l}$ has the second-order belief $b_1^{c_1}$ for which the choice $c_1$ is optimal. If we combine (8.7.25) and (8.7.26) we also conclude that every type $t_2^{c_1,k+l}$ has the second-order belief that assigns probability 1 to $c_2^{k+l-1}$ for which the choice $c_2^{k+l}$ is optimal. These insights, together with (8.7.25) and (8.7.26), imply that every type in the epistemic model constructed above believes in the opponent's rationality. Therefore, every type in the model expresses common belief in rationality.

Now, take an arbitrary choice $c_1 \in C_1^*$. Then, $c_1$ is optimal for the type $t_1^{c_1,k}$ that expresses common belief in rationality. That is, $c_1$ can rationally be chosen under common belief in rationality.

**Case 5.** Suppose that player 1's preferences only depend on his second-order belief and player 2's preferences depend on his first- and second-order belief. Then, every choice $c_1 \in C_1^*$ is optimal for a second-order belief $b_1^{c_1} \in \Delta(C_1^*)$. For every choice $c_1 \in C_1^*$, consider a choice $c_2^1[c_1] \in C_2^*$ for player 2. For every $k \geq 2$ and every $c_1 \in C_1^*$, let $c_2^k[c_1]$ be a choice that is optimal for player 2 if his first-order belief about player 1's choice is given by $b_1^{c_1}$ and believes, with probability 1, that player 1 assigns probability 1 to player 2 choosing $c_2^{k-1}[c_1]$. Since there are finitely many choices, there must be some $k \geq 1$ and $m \geq 1$ such that $c_2^{k+m}[c_1] = c_2^k[c_1]$ for every $c_1 \in C_1^*$.

We construct an epistemic model with sets of types

$$T_1 = \{t_1^{c_1,k+l} \mid c_1 \in C_1^*,\ l \in \{0,...,m-1\}\} \text{ and } T_2 = \{t_2^{c_1,k+l} \mid c_1 \in C_1^*,\ l \in \{0,...,m-1\}\}.$$

The beliefs of the types are as follows:

$$b_1(t_1^{c_1,k+l})(c_2,t_2) := \begin{cases} 1, & \text{if } c_2 = c_2^{k+l}[c_1] \text{ and } t_2 = t_2^{c_1,k+l} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.27}$$

for every $c_1 \in C_1^*$, $l \in \{0,...,m-1\}$ and every $(c_2,t_2) \in C_2 \times T_2$, and

$$b_2(t_2^{c_1,k+l})(c_1',t_1) := \begin{cases} b_1^{c_1}(c_1'), & \text{if } c_1' \in C_1^* \text{ and } t_1 = t_1^{c_1',k+l-1} \\ 0, & \text{otherwise} \end{cases} \tag{8.7.28}$$

for every $c_1 \in C_1^*$, $l \in \{0, ..., m-1\}$ and every $(c_1', t_1) \in C_1 \times T_1$. Here, we use the convention that $t_1^{c_1',k-1} = t_1^{c_1',k+m-1}$.

By combining (8.7.27) and (8.7.28) we see that every type $t_1^{c_1,k+l}$ has the second-order belief $b_1^{c_1}$ for which the choice $c_1$ is optimal. If we combine (8.7.27) and (8.7.28) we also conclude that every type $t_2^{c_1,k+l}$ has the first-order belief $b_1^{c_1} \in \Delta(C_1^*)$, and the second-order belief that assigns probability 1 to $c_2^{k+l-1}[c_1]$, for which the choice $c_2^{k+l}[c_1]$ is optimal. These insights, together with (8.7.27) and (8.7.28), imply that every type in the epistemic model constructed above believes in the opponent's rationality. Therefore, every type in the model expresses common belief in rationality.

Now, take an arbitrary choice $c_1 \in C_1^*$. Then, $c_1$ is optimal for the type $t_1^{c_1,k}$ that expresses common belief in rationality. That is, $c_1$ can rationally be chosen under common belief in rationality.

We have thus explored all 5 cases, which completes the proof. ■

# Solutions to In-Chapter Questions

**Question 8.1.1.** The expected utility of choosing *blue* is $(0.2) \cdot 0 + (0.4) \cdot 6 + (0.4) \cdot 6 = 4.8$, whereas the expected utility of choosing *red* is $(0.2) \cdot 2 + (0.4) \cdot 1 + (0.4) \cdot 0 = 0.8$. Hence, *blue* is the optimal choice.

**Question 8.1.2.** The second-order expectation induced by the upper part is $(0.7) \cdot (b, b) + (0.18) \cdot (r, b) + (0.12) \cdot (r, g)$.

Clearly, the second-order belief in the lower part is different from the upper part: In the upper part, you assign probability 0.3 to the event that Barbara assigns probabilities 0.6 and 0.4 to your choices *blue* and *green,* respectively. In particular, you assign probability 0.3 to the event that Barbara is inherently uncertain about your choice. In contrast, in the lower part you assign probability 0.7 to the event that Barbara assigns probability 1 to your choice *blue,* you assign probability 0.18 to the event that Barbara assigns probability 1 to your choice *blue,* and you assign probability 0.12 to the event that Barbara assigns probability 1 to your choice *green.* In particular, you assign probability 1 to the event that Barbara is certain about your choice, which is fundamentally different from the second-order belief in the upper part.

Nevertheless, the second-order expectation induced by the lower part is also $(0.7) \cdot (b, b) + (0.18) \cdot (r, b) + (0.12) \cdot (r, g)$, like for the upper part.

Given this second-order expectation, the expected utilities induced by your choices are

$$\begin{aligned} u_1(\textit{blue}) &= (0.7) \cdot 0 + (0.18) \cdot 3 + (0.12) \cdot 6 = 1.26, \\ u_1(\textit{green}) &= (0.7) \cdot 4 + (0.18) \cdot 4 + (0.12) \cdot 2 = 3.76, \\ u_1(\textit{red}) &= (0.7) \cdot 2 + (0.18) \cdot 1 + (0.12) \cdot 1 = 1.7, \end{aligned}$$

and hence your optimal choice is *green.*

**Question 8.3.1.** We first derive the second-order expectation that type $t_1^{green}$ has. Note that type $t_1^{green}$ assigns probability 1 to the event that "Barbara chooses *red* and has type $t_2^{red}$ ". In turn, Barbara's type $t_2^{red}$ assigns probability 0.9 to you choosing *blue* and probability 0.1 to you choosing *green.* By putting this together, we see that type $t_1^{green}$ assigns probability 1 to the event that "Barbara chooses *red* and assigns probabilities 0.9 and 0.1 to you choosing *blue* and *green,* respectively".

As such, the second-order expectation $e_1$ induced by type $t_1^{blue}$ is

$$e_1 = (0.9) \cdot (r, b) + (0.1) \cdot (r, g).$$

The expected utilities that you obtain by making your three choices are thus

$$\begin{aligned} u_1(\textit{blue},\, t_1) &= (0.9) \cdot 3 + (0.1) \cdot 6 = 3.3. \\ u_1(\textit{green},\, t_1) &= (0.9) \cdot 4 + (0.1) \cdot 2 = 3.8 \text{ and} \\ u_1(\textit{red},\, t_1) &= (0.9) \cdot 1 + (0.1) \cdot 1 = 1. \end{aligned}$$

As *green* yields the highest expected utility, your choice *green* is optimal for the type $t_1^{green}$.

**Question 8.4.1.** Start with the second-order expectation $(2/3, 0, 0, 1/3)$, where $e_1(w, b) = 2/3$ and $e_1(w, w) = 1/3$, and which assigns probability 0 to the states $(b, w)$ and $(b, b)$. Then,

$$e_1(b_1 \mid w_2) = \frac{e_1(w_2, b_1)}{e_1(w_2, w_1) + e_1(w_2, b_1)} = \frac{2/3}{1/3 + 2/3} = 2/3,$$

which implies that (8.4.5) is satisfied. Moreover, (8.4.6) is trivially satisfied since $e_1$ assigns probability 0 to Barbara choosing *black*.

Consider next the state $(w, b)$, which corresponds to the second-order expectation with $e_1(w, b) = 1$. Then,

$$e_1(b_1 \mid w_2) = \frac{e_1(w_2, b_1)}{e_1(w_2, w_1) + e_1(w_2, b_1)} = \frac{1}{0+1} = 1 \geq 2/3,$$

which implies that (8.4.5) is satisfied. Moreover, (8.4.6) is trivially satisfied since $e_1$ assigns probability 0 to Barbara choosing *black*.

Take now the state $(b, w)$, which corresponds to the second-order expectation with $e_1(b, w) = 1$. Then, (8.4.5) is trivially satisfied since $e_1$ assigns probability 0 to Barbara wearing *white*. Moreover, it also satisfies (8.4.6) as this condition puts no additional restrictions on the second-order expectation.

In a similar way, we can also show that the second-order expectation associated with state $(b, b)$ satisfies the conditions (8.4.5) and (8.4.6).

**Question 8.4.2.** Take a second-order expectation

$$e_1 = \lambda_1 \cdot (2/3, 0, 0, 1/3) + \lambda_2 \cdot (1, 0, 0, 0) + \lambda_3 \cdot (0, 1, 0, 0) + \lambda_4 \cdot (0, 0, 1, 0),$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. We distinguish two cases: (1) $\lambda_1 + \lambda_2 = 0$, and (2) $\lambda_1 + \lambda_2 > 0$.

**Case 1.** Suppose that $\lambda_1 + \lambda_2 = 0$. Thus, $e_1$ assigns probability 0 to Barbara choosing *white,* which means that (8.4.5) is trivially satisfied. As (8.4.6) imposes no additional restrictions, we conclude that (8.4.6) is satisfied as well.

**Case 2.** Suppose that $\lambda_1 + \lambda_2 > 0$. Then, $e_1$ assigns a positive probability to Barbara choosing *white*. Moreover,

$$e_1(w_2, w_1) = \lambda_1 \cdot 1/3 \text{ and } e_1(w_2, b_1) = \lambda_1 \cdot 2/3 + \lambda_2.$$

Thus,

$$\begin{aligned} e_1(b_1 \mid w_2) &= \frac{e_1(w_2, b_1)}{e_1(w_2, w_1) + e_1(w_2, b_1)} = \frac{\lambda_1 \cdot 2/3 + \lambda_2}{\lambda_1 \cdot 1/3 + (\lambda_1 \cdot 2/3 + \lambda_2)} \\ &= \frac{\lambda_1 \cdot 2/3 + \lambda_2}{\lambda_1 + \lambda_2} = \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot 2/3 + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot 1 \geq 2/3. \end{aligned}$$

This implies that (8.4.5) is satisfied. As (8.4.6) imposes no additional restrictions, condition (8.4.6) is satisfied as well.

**Question 8.4.3.** Consider the sets of types $T_1 = \{t_1^{black}, \hat{t}_1^{black}, t_1^{white}\}$ and $T_2 = \{t_2^{white}, \hat{t}_2^{white}, t_2^{black}\}$ where

$$\begin{aligned} b_1(t_1^{black}) &= (\textit{white}, t_2^{white}), \\ b_1(\hat{t}_1^{black}) &= (0.5) \cdot (\textit{white}, \hat{t}_2^{white}) + (0.5) \cdot (\textit{black}, t_2^{black}), \\ b_1(t_1^{white}) &= (\textit{black}, t_2^{black}), \\ b_2(t_2^{white}) &= (0.6) \cdot (\textit{black}, \hat{t}_1^{black}) + (0.4) \cdot (\textit{white}, t_1^{white}), \\ b_2(\hat{t}_2^{white}) &= (0.5) \cdot (\textit{black}, \hat{t}_1^{black}) + (0.5) \cdot (\textit{white}, t_1^{white}), \\ b_2(t_2^{black}) &= (\textit{white}, t_1^{white}). \end{aligned}$$

**Question 8.5.1. (a)** The beliefs in the partial beliefs diagram induce the second-order expectation

$$e_1 = (0.5) \cdot (7, 1) + (0.5) \cdot (3, 7).$$

**(b)** If you practice for 5 weeks, then your utility under the two states $(7, 1)$ and $(3, 7)$ are given by

$$u_1(5, (7, 1)) = 5 \cdot 7 - 5^2 + (5 - 1) = 14$$

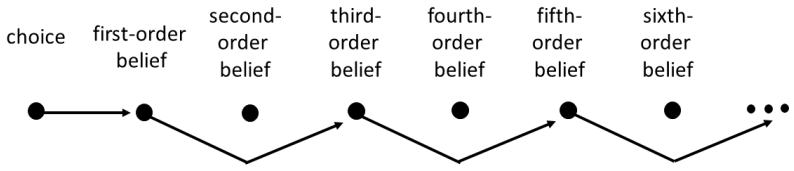
and

$$u_1(5, (3, 7)) = 5 \cdot 3 - 5^2 = -10.$$

Note that in the second utility there is no mental bonus for exceeding Barbara's expectations, as your choice 5 would be below Barbara's expectation 7. The expected utility is thus given by

$$u_1(5, e_1) = (0.5) \cdot 14 + (0.5) \cdot (-10) = 2.$$

**Question 8.6.1.** The causality diagram that belongs to this game is the following:




Indeed, to support a given choice for you we only need a first-order belief. In turn, to justify this first-order belief under common belief in rationality, we need to specify what you believe about Barbara's second-order belief. That is, we need to specify your third-order belief. To support this third-order belief under common belief in rationality we need to specify what you believe that Barbara believes that you believe about Barbara's second-order belief. In other words, we must specify your fifth-order belief. And so on.

# Problems

**Problem 8.1: The hieroglyphs exam.**

You and Barbara are both fascinated by ancient Egypt, and have recently started to study the ancient egyptian hieroglyphs. Next month you will both take an exam on hieroglyphs at the Open University, and you must decide which grade you would like to obtain for the exam. This is not an easy problem since you do not only care about your grade, but also about (i) the time and effort it takes to obtain that grade, (ii) how close your grade is to Barbara's, and (iii) whether you will be able to exceed Barbara's expectations by your grade.

More precisely, raising the grade by 1 point will on the one hand increase your utility by 3 units because you enjoy writing a good exam. On the other hand, achieving a grade of $g_1$ will lower your utility by $g_1^2$ units because of the time and effort you would need to put in. Moreover, since Barbara is such a good friend, you would like to have your grade close to Barbara's grade: Every point by which your grades differ will lower your utility by an additional 5 units. Finally, you would like to exceed Barbara's expectations if possible: If your grade is higher than what you believe Barbara expects, then this will increase your utility by an additional 10 units.

All this can be represented by your utility function $u_1$, where

$$u_1(g_1, (g_2, g_1')) = 3g_1 - g_1^2 - 5 \cdot |g_1 - g_2| \; + \begin{cases} 10, & \text{if } g_1 > g_1' \\ 0, & \text{otherwise} \end{cases} .$$

Here, $g_1$ is the grade you would like to achieve, $g_2$ is the grade you believe that Barbara wants to achieve, and $g_1'$ is the grade that you believe that Barbara believes that you try to achieve. Barbara's utility function is similar.

To keep things easy, assume that you only consider achieving a 6 (out of 10), a 7, an 8 or a 9 as your grade, and similarly for Barbara.

**(a)** Translate this story into a psychological game, by specifying your decision problem. The decision problem for Barbara is the same, by symmetry.

**(b)** Suppose you would like to find the choices you can rationally make under common belief in rationality. Is the iterated elimination of choices and states guaranteed to be sufficient for this purpose? Explain your answer.

**(c)** Find the grades that you can rationally try to achieve under common belief in rationality. Which procedure do you use?

**(d)** Based on the outcome in (c), create a beliefs diagram with solid arrows only that uses all the choices for you and Barbara that can rationally be made under common belief in rationality. Translate the beliefs diagram into an epistemic model where all types express common belief in rationality.

**Problem 8.2: Time to choose a sport.**

You and Barbara have been studying a lot for the hieroglyphs exam lately. A bit too much, actually, since you both have not done any physical exercise during this period. To compensate for this, you both want to start doing some sports. In your neighbourhood there is the choice between *football, rugby, handball, water polo, swimming* and *athletics.* Today, you and Barbara must both choose one of these sports.

| | *football* | *rugby* | *handball* | *water polo* | *swimming* | *athletics* |
|---|---|---|---|---|---|---|
| You | 4 | 16 | 8 | 24 | 20 | 12 |
| Barbara | 12 | 20 | 24 | 8 | 4 | 16 |

Table 8.7.1 Baseline utilities in Problem 8.2

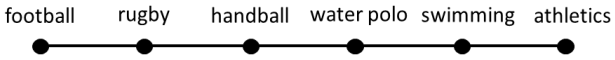



Figure 8.7.1 Differences between the sports in Problem 8.2

When making your choice, you do not only care about how much you enjoy that sport, but you would also like to *surprise* Barbara as much as possible by your choice. Barbara, on the other hand, cares about choosing a sport that is as *different* as possible from the sport that you choose.

More precisely, the baseline utilities you and Barbara derive from choosing one of these sports are given by Table 8.7.1. These utilities indicate how much you and Barbara like that particular sport.

The differences between the various sports can be visualized by Figure 8.7.1. Assume that the distance between two neighbouring dots is always 1. Then, the difference between two sports is measured by the distance between their respective dots in the figure. For instance, the difference between *rugby* and *water polo* is 2, whereas the distance between *football* and *swimming* is 4. For two sports $s$ and $s'$ we denote their difference by $d(s, s')$.

Suppose that you choose the sport $s_1$ and that you believe that Barbara believes that you choose the sport $s_1'$. Then, your *surprise utility* is given by

$$(d(s_1, s_1'))^2.$$

Hence, the higher the difference between your actual choice and what you believe that Barbara believes that you choose, the higher your surprise utility. Moreover, larger differences have a higher impact on your preferences than smaller differences. Your total utility is the sum of your baseline utility and your surprise utility.

Similarly, if you choose the sport $s_1$ and Barbara chooses the sport $s_2$, then Barbara's *utility from being different* is given by

$$(d(s_1, s_2))^2.$$

Barbara's total utility is the sum of her baseline utility and her utility from being different.

The question is: Which sport will you choose?

**(a)** Translate this story into a psychological game, by specifying the decision problems for you and Barbara.

**(b)** Suppose you would like to find the choices you can rationally make under common belief in rationality. Is the iterated elimination of choices and states guaranteed to be sufficient for this purpose? Explain your answer.

**(c)** Find the sports you can rationally choose under common belief in rationality. Which procedure do you use?

**(d)** Based on the outcome in (c), create a beliefs diagram with solid arrows only that uses all the sports that you can rationally choose under common belief in rationality. Translate the beliefs diagram into an epistemic model where all types express common belief in rationality.

### *Problem 8.3: How to disappoint Barbara?

Chris is moving to a smaller apartment, and wants to get rid of all the things he has not used during the last few years. He just came across a Beatles-mug that he received as a present from a friend many years ago, but has never used. Barbara and you, however, are both lifelong Beatles fans and would really like to have the mug. Since there is only one Beatles-mug, Chris decides to auction it. The rules of the auction are as follows: You and Barbara must simultaneously whisper a price in Chris' ear, which must be either 20 or 40 euros, and the person who names the highest price will get the mug and must pay the price he or she chose. If you both name the same price, Chris will toss a coin to decide who gets the mug.

You and Barbara both value the mug at 30 euros. Consequently, your baseline utility if you win the auction will be $30 - p$, where $p$ is the price you pay. If you lose the auction your baseline utility will be 0.

Yesterday, you and Barbara entered into a long fight about who was the best musician among the four Beatles, and for that reason you would like to disappoint Barbara if possible. More precisely, if you happen to win the auction, but believe that Barbara believes that she would win the auction with probability 1, then you believe that Barbara will be very disappointed. In that case, your utility will be the baseline utility plus a disappointment bonus of 160. In all other cases, your utility will just be the baseline utility. The utilities for Barbara are similar.

**(a)** Translate this story into a psychological game, by specifying the decision problem for you. The decision problem for Barbara will be similar.

**(b)** Suppose you would like to find the choices you can rationally make under common belief in rationality. Is the iterated elimination of choices and states guaranteed to be sufficient for this purpose? Explain your answer.

**(c)** Apply the first four rounds of the *iterated elimination of choices and second-order expectations* to find those second-order expectations for you that are consistent with up to three-fold belief in rationality.

It turns out that the second-order expectations consistent with *common* belief in rationality are approximately the second-order expectations you found in (c).

**(d)** Which prices can you rationally whisper into Chris' ear under common belief in rationality?

**(e)** Based on the set of second-order expectations found in (c), and using the method outlined in Section 8.4.9, create a beliefs diagram with solid arrows only that involves all the prices you found in (d). Translate this beliefs diagram into an epistemic model where all types express common belief in rationality.

# Literature

**Psychological games.** The first to introduce psychological games were Geanakoplos, Pearce and Stacchetti (1989). In their model, the utility of a player may depend on his choice, and – potentially – on *all* layers of his belief hierarchy. It is thus more general than our setting, where we assume that the utility of a player depends at most on the first two layers of his belief hierarchy, that is, his first- and second-order belief. Moreover, we make the additional assumptions that the utility only depends on the *second-order expectation*, which is a summary statistic of the first- and second-order belief, and that the utility depends *linearly* on the second-order expectation. Indeed, we assume that the conditional preferences of the player have an expected utility representation, which is equivalent to saying that the player's utility depends linearly on his second-order expectation. The model by Geanakoplos, Pearce and Stacchetti (1989), in contrast, does not make any of these assumptions, and is therefore also more general along those dimensions.

In addition, the model of Geanakoplos, Pearce and Stacchetti (1989) also covers dynamic games, whereas we restrict attention to static games in this book. In dynamic games, their model specifies that the utility of a player depends on his hierarchy of *initial* beliefs and the *outcome* of the game. Later, Battigalli and Dufwenberg (2009) extended their model by (i) allowing the player's utility to depend on *conditional* beliefs during the game, and not only on *initial* beliefs at the beginning, and (ii) allowing the player's utility to depend on the full *strategies* of the players, and not only on the *outcome* that is finally reached. See also Battigalli, Corrao and Dufwenberg (2019).

**Linear psychological games.** As already mentioned, the model in this chapter is more restrictive than the original model in Geanakoplos, Pearce and Stacchetti (1989), because we assume that (i) the player's utility only depends on his *first-* and *second-order* belief, (ii) in fact, the player's utility only depends on the *second-order expectation* induced by his first- and second-order belief, and (iii) the player's utility depends *linearly* on his second-order expectation.

As such, the model we use belongs to the class of belief-finite *linear* psychological games as introduced by Jagau and Perea (2023). In a belief-finite linear psychological game, (i) the utility of a player only depends on finitely many layers of beliefs, that is, only depends on his beliefs up to order $n$, for some $n$, (ii) the utility only depends on the $n$-th order expectation induced by his beliefs up to order $n$, and (iii) the utility is linear in his $n$-th order expectation. Hence, in this chapter we focus on belief-finite linear psychological games where $n$ is equal to 2. Also, the concept of *higher-order expectation* is taken from Jagau and Perea (2023).

**Procedure for common belief in rationality.** In this chapter we have introduced a recursive elimination procedure, called the *iterated elimination of choices and second-order expectations,* which yields precisely those choices that are possible under common belief in rationality. This procedure is based on the *iterated elimination of choices and n-th order beliefs* in Jagau and Perea (2022) for psychological games that are not necessarily linear, and the *iterated elimination of choices and n-th order expectations* in Jagau and Perea (2023) for psychological games that are linear. A related procedure can be found in Battigalli, Corrao and Sanna (2020) which characterizes the strategies that can rationally be made under common *strong* belief in rationality in a *dynamic* psychological game.

**(Im)possibility of common belief in rationality.** In a psychological game with finitely many choices, but where a player's utility depends on *all* layers of his belief hierarchy, it may happen that there is *no* belief hierarchy that expresses *common belief in rationality.* Examples can be found in Bjorndahl, Halpern and Pass (2013) and Jagau and Perea (2022). Both papers provide sufficient

conditions on the players' utility functions which guarantee that for every player there *is* a belief hierarchy that expresses common belief in rationality.

**When elimination of choices and states is enough.** In Section 8.6 we have investigated for which classes of psychological games the *iterated elimination of choices and states* is sufficient to find all choices that can rationally be made under common belief in rationality. The concept of a *causality diagram* played a crucial role in that analysis. Our findings in Theorem 8.6.1, as well as the notion of a causality diagram, are based on Mourmans (2019). He explored this issue for the class of *all* belief-finite linear psychological games, including psychological games where the player's utility may depend on his *third-order* belief or *higher*. Mourmans (2019) showed a little more in his theorem than we did in our Theorem 8.6.1: He did not only identify the classes of belief-finite linear psychological games where the iterated elimination of choices and states is sufficient for finding those choices that are possible under common belief in rationality, he also showed that for every type of psychological game *outside* this class we can always find associated utility functions for the players such that the iterated elimination of choices and states *fails* to identify precisely those choices that are possible under common belief in rationality.