Chapter 9

Correct and Symmetric Beliefs in Psychological Games

In Chapter 4 we have formalized the ideas of *correct* and *symmetric* beliefs for standard games by means of *simple* and *symmetric belief hierarchies,* respectively. It turns out that these two notions can be applied without any change to the class of psychological games. We show that common belief in rationality in combination with a simple belief hierarchy leads to the concept of *psychological Nash equilibrium,* whereas combining common belief in rationality with a symmetric belief hierarchy yields the concept of *psychological correlated equilibrium.* Similarly to Chapter 4 we will see that every psychological game has at least one psychological Nash equilibrium and at least one psychological correlated equilibrium. As a consequence, combining the conditions of common belief in rationality with those of a simple, or symmetric, belief hierarchy never leads to logical contradictions. In Chapter 9 of the online appendix we discuss some economic applications.

9.1 Correct Beliefs

In Chapters 4 and 6 we have explored the idea of *correct beliefs,* which states that you believe that your opponent is correct about the beliefs you hold. We have seen that this idea can be formalized by the notion of a *simple belief hierarchy.* In a standard game, a simple belief hierarchy is fully generated by a single belief about player 1's choice, a single belief about player 2's choice, and so on. This definition of a simple belief hierarchy can be carried over *without any change* to psychological games.

We will see that if we combine the conditions of common belief in rationality with those of a simple belief hierarchy, then we obtain the concept of *psychological Nash equilibrium.* This is similar to what we have seen in Chapter 4, where common belief in rationality in combination with a simple belief hierarchy led to *Nash equilibrium* in standard games, and also similar to Chapter 6 which revealed that in games with incomplete information, common belief in rationality together with a simple belief

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 9.1.1 Decision problems for "Barbara's birthday"

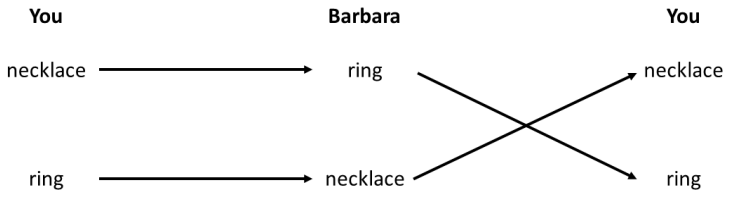

Figure 9.1.1 Beliefs diagram for "Barbara's birthday"

hierarchy yields the notion of *generalized Nash equilibrium.*

9.1.1 Simple Belief Hierarchies

To illustrate the idea of correct beliefs in psychological games, let us go back to the example "Barbara's birthday" which we explored already in Section 8.6.

Example 9.1: Barbara's birthday.

Recall the story from Section 8.6. For convenience, we have reproduced the decision problems for you and Barbara in Table 9.1.1. Recall that $(\cdot, n)$ represents the collection of states in your decision problem where you believe that Barbara believes that you buy a *necklace,* and that $(n, \cdot)$ is the collection of states in Barbara's decision problem where Barbara believes that you buy a *necklace.* Similarly for the other collections of states in the two decision problems.

We have seen in Section 8.6 that under common belief in rationality you can rationally buy a *necklace* or a *ring.* This insight is supported by the beliefs diagram in Figure 9.1.1. Indeed, the belief hierarchies that support your choices *necklace* and *ring* both express common belief in rationality. In particular, under common belief in rationality you can rationally buy a *necklace* if you believe, with probability 1, that Barbara believes, with probability 1, that you will buy a *ring.* In this case, you believe to *surprise Barbara with probability* 1 by buying a *necklace.* Similarly, under common belief in rationality you can also believe to surprise Barbara with probability 1 by buying a *ring.*

However, in both belief hierarchies you believe that Barbara is *incorrect* about your beliefs. Consider, for instance, the belief hierarchy that supports your choice *necklace.* There, you believe that Barbara believes that you will buy a *ring,* but at the same time you believe that Barbara believes that you believe that Barbara believes that you will buy a *necklace.* That is, you believe that Barbara is wrong about your second-order belief. Or, equivalently, your second-order belief and fourth-order belief do not correspond to a single probabilistic belief σ_1 about your choice.

Recall from Chapter 4 that in a two-player standard game, all higher-order beliefs in a *simple hierarchy* are generated by a *single* belief σ_1 about player 1's choice, and a *single* belief σ_2 about

player 2's choice. This will also be precisely the definition of a simple belief hierarchy for psychological games.

Definition 9.1.1 (Simple belief hierarchy) *Let σ_1 be a probabilistic belief about player 1's choice and σ_2 a probabilistic belief about player 2's choice. The belief hierarchy for player i generated by the beliefs (σ_1, σ_2) is defined as follows:*

(1) in the first-order belief, player i assigns to every opponent's choice c_j the probability $\sigma_j(c_j)$,

(2) in the second-order belief, player i believes with probability 1 that opponent j assigns to every choice c_i for player i the probability $\sigma_i(c_i)$,

(3) in the third-order belief, player i believes with probability 1 that opponent j believes with probability 1 that player i assigns to every opponent's choice c_j the probability $\sigma_j(c_j)$, and so on.

*A belief hierarchy is called **simple** if it is generated by a pair of such beliefs (σ_1, σ_2).*

In order words, σ_j generates i's first-order belief, third-order belief, and so on, whereas σ_i generates i's second-order belief, fourth-order belief, and so on.

In the beliefs diagram of Figure 9.1.1 it can thus be verified that none of your belief hierarchies is simple, and that none of Barbara's belief hierarchies is simple. Recall that in both of your belief hierarchies, you believe you are able to surprise Barbara *with probability* 1. However, later in this section we will see that if you hold a *simple* belief hierarchy that expresses common belief in rationality, then you believe that you will only be able to surprise Barbara *with probability at most* 0.6.

9.1.2 Relation with Psychological Nash Equilibrium

Suppose we combine the property of a simple belief hierarchy with the conditions of common belief in rationality. What can we say about the belief hierarchy in this case? That is the question we will focus on next.

Consider a simple belief hierarchy for player i generated by the pair of beliefs (σ_1, σ_2). If player i believes in j's rationality, then in this belief hierarchy player i must only assign positive probability to opponent j's choices c_j which are optimal for player j, given what i believes about j's second-order expectation.

By construction, i's belief about j's choice is given by the belief σ_j. Hence, the opponent's choices to which i assigns positive probability are precisely the choices c_j with $\sigma_j(c_j) > 0$.

But what does player i believe about j's second-order expectation? By definition, i believes that j's first-order belief about i's choice is given by σ_i. Moreover, i believes that j's second-order belief about the belief that i has about j's choice is given by σ_j. Put together, i believes that j's second-order expectation assigns to every pair $(c_i, c_j) \in C_i \times C_j$ the probability

$$\sigma_i(c_i) \cdot \sigma_j(c_j).$$

We say that this is j's second-order expectation *induced* by (σ_1, σ_2). Here is the formal definition.

Definition 9.1.2 (Induced second-order expectation) *Consider a pair of beliefs (σ_1, σ_2), where σ_1 is a probabilistic belief about 1's choice, and σ_2 is a probabilistic belief about 2's choice. For both players i, the **second-order expectation** $e_i[\sigma_1, \sigma_2]$ **induced by** (σ_1, σ_2) is the probability distribution that assigns to every pair of choices $(c_j, c_i) \in C_j \times C_i$ the probability*

$$\sigma_j(c_j) \cdot \sigma_i(c_i).$$

Question 9.1.1 *In the example "Barbara's birthday", consider the pair of beliefs* (σ_1, σ_2) *where*

$$\sigma_1 = (0.2) \cdot \textit{ necklace } + (0.5) \cdot \textit{ ring } + (0.3) \cdot \textit{ bracelet}$$

and

$$\sigma_2 = (0.6) \cdot \textit{ necklace } + (0.4) \cdot \textit{ bracelet.}$$

Here, you are player 1 and Barbara is player 2. Find the induced second-order expectation $e_1[\sigma_1, \sigma_2]$ *for you and the induced second-order expectation* $e_2[\sigma_1, \sigma_2]$ *for Barbara.*

Recall from above that if player i believes in j's rationality, then i must only assign positive probability to opponent j's choices c_j which are optimal for player j, given what i believes about j's second-order expectation. Combining this with the insights above, we conclude for every opponent's choice c_j that

$$\sigma_j(c_j) > 0 \text{ only if } c_j \text{ is optimal for the induced second-order expectation } e_j[\sigma_1, \sigma_2]. \tag{9.1.1}$$

Suppose now that, in addition, player i also believes that j believes in i's rationality. Then, i believes that j will only assign positive probability to choices c_i for player i that are optimal, given what i believes that j believes about i's second-order expectation. By construction, i's belief about j's belief about i's choice is σ_i. Moreover, i believes that j believes that i has the second-order expectation $e_i[\sigma_1, \sigma_2]$ induced by (σ_1, σ_2). Putting all these insights together, we conclude that

$$\sigma_i(c_i) > 0 \text{ only if } c_i \text{ is optimal for the induced second-order expectation } e_i[\sigma_1, \sigma_2]. \tag{9.1.2}$$

Hence, if player i's simple belief hierarchy is generated by the pair of beliefs (σ_1, σ_2), player i believes in j's rationality, and believes that j believes in i's rationality, then the pair of beliefs (σ_1, σ_2) must satisfy the properties (9.1.1) and (9.1.2) above. Such pairs of beliefs are called *psychological Nash equilibria.*

Definition 9.1.3 (Psychological Nash equilibrium) *Consider a probabilistic belief* σ_1 *about player 1's choice and a probabilistic belief* σ_2 *about player 2's choice. The pair of beliefs* (σ_1, σ_2) *is a* ***psychological Nash equilibrium*** *if for both players* i*, and for every choice* $c_i \in C_i$*, we have that*

$$\sigma_i(c_i) > 0 \textit{ only if } c_i \textit{ is optimal for the induced second-order expectation } e_i[\sigma_1, \sigma_2].$$

We thus see that if player i's simple belief hierarchy is generated by the pair of beliefs (σ_1, σ_2), and player i expresses (the first two layers of) common belief in rationality, then (σ_1, σ_2) must be a psychological Nash equilibrium.

We now show that the other direction is also true: If (σ_1, σ_2) is a psychological Nash equilibrium, then the simple belief hierarchy generated by it will express common belief in rationality. To see why, consider a psychological Nash equilibrium (σ_1, σ_2) and the simple belief hierarchy for player i generated by it. To show that i believes in j's rationality, suppose that i assigns a positive probability to choice c_j. As i's belief about j's choice is given by σ_j, it must be that $\sigma_j(c_j) > 0$. By the definition of a psychological Nash equilibrium, it must then be that c_j is optimal for the induced second-order expectation $e_j[\sigma_1, \sigma_2]$. As i's simple belief hierarchy is generated by (σ_1, σ_2), player i believes that j's second-order expectation is given by $e_j[\sigma_1, \sigma_2]$. Hence, i only assigns a positive probability to j's choice c_j if c_j is optimal for j, given what i believes about j's second-order expectation. In other words, i believes in j's rationality.

Question 9.1.2 *Explain, by a similar argument, that i also believes that j believes in i's rationality.*

If we continue in this fashion, we also conclude that player i expresses 3-fold belief in rationality, 4-fold belief in rationality, and so on, *ad infinitum.* That is, player i's belief hierarchy expresses common belief in rationality. By combining all the insights above we arrive at the following general conclusion.

Theorem 9.1.1 (Relation with psychological Nash equilibrium) *Consider the simple belief hierarchy for player i generated by a belief pair (σ_1, σ_2). Then, this belief hierarchy expresses common belief in rationality, if and only if, the belief pair (σ_1, σ_2) is a psychological Nash equilibrium.*

In other words, combining the condition of a simple belief hierarchy with the conditions in common belief in rationality yields precisely the concept of *psychological Nash equilibrium.* Observe the similarity with Theorem 4.1.1, which states that in a standard game the combination of common belief in rationality with a simple belief hierarchy leads to *Nash equilibrium.* Or compare it with Theorem 6.1.1, which states that in a game with incomplete information the same conditions lead to *generalized Nash equilibrium.*

We eventually want to characterize the *choices* that player i can rationally make if he holds a simple belief hierarchy that expresses common belief in rationality. Theorem 9.1.1 is the key to completing this task. Suppose that player i holds a simple belief hierarchy which is generated by the belief pair (σ_1, σ_2) and expresses common belief in rationality, and assume that the choice c_i is optimal for this belief hierarchy. Then, the choice c_i is optimal for the second-order expectation that player i holds in this belief hierarchy. By construction, this second-order expectation is $e_i[\sigma_1, \sigma_2]$ – the second-order expectation induced by (σ_1, σ_2). Since the simple belief hierarchy expresses common belief in rationality, we know by Theorem 9.1.1 that the belief pair (σ_1, σ_2) is a psychological Nash equilibrium. Altogether, we thus see that every choice which is optimal for a simple belief hierarchy that expresses common belief in rationality must be optimal for the second-order expectation induced by a psychological Nash equilibrium.

We now show that the other direction is also true: Every choice that is optimal for the second-order expectation induced by a psychological Nash equilibrium is optimal for a simple belief hierarchy that expresses common belief in rationality. To see this, consider a choice c_i that is optimal for the second-order expectation $e_i[\sigma_1, \sigma_2]$ induced by a psychological Nash equilibrium (σ_1, σ_2). Then, the choice c_i is optimal for the simple belief hierarchy generated by (σ_1, σ_2). Since (σ_1, σ_2) is a psychological Nash equilibrium, we know by Theorem 9.1.1 that the simple belief hierarchy generated by (σ_1, σ_2) expresses common belief in rationality. Hence, the choice c_i is optimal for a simple belief hierarchy that expresses common belief in rationality.

By combining the two insights above, we reach the following conclusion.

Theorem 9.1.2 (Relation with psychological Nash equilibrium choices) *A choice is optimal for a simple belief hierarchy that expresses common belief in rationality, if and only if, that choice is optimal for the second-order expectation induced by a psychological Nash equilibrium.*

This result has great practical value: Indeed, if we wish to find all choices that can rationally be made with a simple belief hierarchy that expresses common belief in rationality, then it suffices to find all psychological Nash equilibria in the game, and subsequently determine which choices are optimal in these psychological Nash equilibria.

9.1.3 Examples

We will now return to two examples we have introduced in Chapter 8, and use Theorems 9.1.1 and 9.1.2 to find all the simple belief hierarchies that express common belief in rationality, and all the choices you can rationally make with such belief hierarchies.

Example 9.2: Barbara's birthday.

Recall the decision problems in Table 9.1.1. We first wish to find all simple belief hierarchies for you that express common belief in rationality. By Theorem 9.1.1 these are precisely the simple belief hierarchies generated by a psychological Nash equilibrium (σ_1, σ_2). Hence, our first task is to find all psychological Nash equilibria (σ_1, σ_2) in the game.

Question 9.1.3 *Explain why in a psychological Nash equilibrium (σ_1, σ_2), the belief σ_1 should assign probability zero to your choice bracelet, and the belief σ_2 should assign probability zero to Barbara's choice bracelet.*

Let (σ_1, σ_2) be a psychological Nash equilibrium in the game. In view of Question 9.1.3 we know that $\sigma_1(\textit{bracelet}) = 0$ and $\sigma_2(\textit{bracelet}) = 0$. We distinguish two cases: (1) $\sigma_1(\textit{necklace}) > 0$ and (2) $\sigma_1(\textit{ring}) > 0$.

Case 1. Suppose that $\sigma_1(\textit{necklace}) > 0$. Then, *necklace* must be optimal for you under the induced second-order expectation $e_1[\sigma_1, \sigma_2]$. Note that your conditional preference relation only depends on your second-order belief, which is given by σ_1. Hence, your choice *necklace* must be optimal under your second-order belief σ_1. In view of your decision problem in Table 9.1.1, and given the fact that $\sigma_1(\textit{bracelet}) = 0$, this is only possible if $\sigma_1(\textit{ring}) > 0$. We thus see that $\sigma_1(\textit{necklace}) > 0$ and $\sigma_1(\textit{ring}) > 0$.

By definition of a psychological Nash equilibrium, this means that both *necklace* and *ring* must be optimal for you under the same second-order belief σ_1. In particular, the expected utilities for *necklace* and *ring* must be the same under the second-order belief σ_1. If we look at your decision problem in Table 9.1.1, these expected utilities are given by

$$u_1(\textit{necklace},\ \sigma_1) = \sigma_1(\textit{ring}) \cdot 3 \text{ and } u_1(\textit{ring},\ \sigma_1) = \sigma_1(\textit{necklace}) \cdot 2.$$

As both expected utilities must be equal, we have that

$$3 \cdot \sigma_1(\textit{ring}) = 2 \cdot \sigma_1(\textit{necklace}).$$

Since $\sigma_1(\textit{necklace}) = 1 - \sigma_1(\textit{ring})$, we conclude that

$$3 \cdot \sigma_1(\textit{ring}) = 2 \cdot (1 - \sigma_1(\textit{ring})),$$

and hence

$$\sigma_1(\textit{ring}) = 0.4,$$

which implies that

$$\sigma_1(\textit{necklace}) = 0.6.$$

We thus see that

$$\sigma_1 = (0.6) \cdot \ \textit{necklace} + (0.4) \cdot \textit{ring}.$$

As Barbara's preferences only depend on her first-order belief σ_1, we conclude that Barbara's unique optimal choice is *necklace.* By definition of a psychological Nash equilibrium, the belief σ_2 should therefore assign probability 1 to Barbara's choice *necklace.*

Altogether, we see that in Case 1 the only psychological Nash equilibrium (σ_1, σ_2) is given by

$$\sigma_1 = (0.6) \cdot \textit{ necklace} + (0.4) \cdot \textit{ring} \text{ and } \sigma_2 = \textit{ necklace.}$$

Case 2. Suppose that $\sigma_1(\textit{ring}) > 0$. Then, *ring* must be optimal for you under your second-order belief σ_1. In view of your decision problem in Table 9.1.1, and given the fact that $\sigma_1(\textit{bracelet}) = 0$, this is only possible if $\sigma_1(\textit{necklace}) > 0$. We thus see that $\sigma_1(\textit{necklace}) > 0$ and $\sigma_1(\textit{ring}) > 0$. But this situation has been covered in Case 1 already, and led to the psychological Nash equilibrium above.

Hence, there is a *unique* psychological Nash equilibrium in this game, given by

$$\sigma_1 = (0.6) \cdot \textit{ necklace} + (0.4) \cdot \textit{ring} \text{ and } \sigma_2 = \textit{ necklace.}$$

In this psychological Nash equilibrium, you are indifferent between the choices *necklace* and *ring,* whereas Barbara's unique optimal choice is *necklace.* In view of Theorem 9.1.2 we thus conclude that with a simple belief hierarchy that expresses common belief in rationality, you can rationally buy a *necklace* or a *ring,* whereas Barbara can only rationally guess that you buy a *necklace.*

But we can say a bit more: Recall from above that under common belief in rationality, but without insisting on a simple belief hierarchy, you can believe to surprise Barbara with *probability* 1 if you buy a *necklace* or a *ring.* However, if we additionally require you to hold a *simple* belief hierarchy, then you can only believe to surprise Barbara with *probability at most* 0.6. To see this, note that under common belief in rationality with a simple belief hierarchy your second-order belief must be

$$(0.6) \cdot \textit{ necklace} + (0.4) \cdot \textit{ring.}$$

That is, you must believe that Barbara assigns probabilities 0.6 and 0.4 to your choices *necklace* and *ring,* respectively. But then, by buying a *necklace* you believe to surprise Barbara only with probability 0.4, whereas by buying a *ring* you believe to surprise Barbara with probability 0.6. Hence, you believe to surprise Barbara with probability 0.6 at most.

There is also a clear intuition for this phenomenon: By imposing a simple belief hierarchy, we require you to believe that Barbara is *correct* about your second-order belief. This, in turn, heavily restricts your possibilities of surprising Barbara.

Indeed, full surprise by buying a *ring,* for instance, would only be possible if you believe with probability 1 that Barbara believes with probability 1 that you make the other choice *necklace.* However, if you believe that Barbara is correct about your second-order belief, and believe that Barbara believes in your rationality, then you must believe that Barbara is correct about your actual choice *ring* as well. But then, choosing *ring* would no longer be optimal, since you would not be able to surprise Barbara by it. A similar argument tells us that full surprise by buying a *necklace* is not possible either.

Therefore, full surprise is ruled out if we insist on a simple belief hierarchy. In fact, it turns out that under common belief in rationality with a simple belief hierarchy, surprising Barbara is only possible up to a degree of 0.6.

Example 9.3: Surprising Barbara.

Recall the story from Section 8.1. For convenience, we reproduce the decision problems for you and Barbara in Table 9.1.2. Note that your choice *red* is strictly dominated by the randomized choice

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 9.1.2 Decision problems for "Surprising Barbara"

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 9.1.3 Reduced decision problems for "Surprising Barbara"

$(0.4)\cdot$ *blue* $+ (0.6)\cdot$ *green,* and that Barbara's choice *green* is strictly dominated by $(0.6)\cdot$ *blue* $+ (0.4)\cdot$ *red.* Therefore, your choice *red* and Barbara's choice *green* are never optimal for any second-order expectation. As a consequence, every belief hierarchy that expresses common belief in rationality must assign, at each of its layers, probability zero to your choice *red* and Barbara's choice *green.* We may thus restrict ourselves to the reduced decision problems in Table 9.1.3.

We will first try to find all simple belief hierarchies for you that express common belief in rationality. Consider a simple belief hierarchy for you generated by the pair of beliefs (σ_1, σ_2). If this belief hierarchy expresses common belief in rationality, we know from above that it must assign probability zero to your choice *red* and to Barbara's choice *green* at each of its layers. This means that σ_1 must assign probability zero to your choice *red,* and that σ_2 must assign probability zero to Barbara's choice *green.* That is, (σ_1, σ_2) operates entirely within the reduced decision problems of Table 9.1.3.

Moreover, we know from Theorem 9.1.1 that (σ_1, σ_2) must be a psychological Nash equilibrium. We distinguish two cases: (1) $\sigma_1(\textit{blue}) > 0$, and (2) $\sigma_1(\textit{blue}) = 0$.

Case 1. Suppose that $\sigma_1(\textit{blue}) > 0$. Then, by definition of a psychological Nash equilibrium, your choice *blue* must be optimal under the induced second-order expectation $e_1[\sigma_1, \sigma_2]$. In view of your decision problem in Table 9.1.3, this is only possible if $\sigma_1(\textit{green}) > 0$. We thus see that $\sigma_1(\textit{blue}) > 0$ and $\sigma_1(\textit{green}) > 0$, which means that both *blue* and *green* must be optimal for you under the same second-order expectation $e_1[\sigma_1, \sigma_2]$. In particular, your choices *blue* and *green* must yield the same expected utility under $e_1[\sigma_1, \sigma_2]$. These expected utilities are equal to

$$u_1(\textit{blue},\, e_1[\sigma_1, \sigma_2]) = \sigma_2(b) \cdot \sigma_1(g) \cdot 3 + \sigma_2(r) \cdot \sigma_1(b) \cdot 3 + \sigma_2(r) \cdot \sigma_1(g) \cdot 6$$

and

$$u_1(\textit{green},\, e_1[\sigma_1, \sigma_2]) = \sigma_2(b) \cdot \sigma_1(b) \cdot 4 + \sigma_2(b) \cdot \sigma_1(g) \cdot 2 + \sigma_2(r) \cdot \sigma_1(b) \cdot 4 + \sigma_2(r) \cdot \sigma_1(g) \cdot 2.$$

Setting these two expected utilities equal leads to

$$\sigma_1(b) \cdot [4\sigma_2(b) + \sigma_2(r)] = \sigma_1(g) \cdot [\sigma_2(b) + 4\sigma_2(r)]. \tag{9.1.3}$$

We distinguish two subcases: (1.1) $\sigma_2(\textit{blue}) > 0$ and (1.2) $\sigma_2(\textit{blue}) = 0$.

Case 1.1. Suppose that $\sigma_2(\textit{blue}) > 0$. Then, *blue* must be optimal for Barbara under the induced second-order expectation $e_2[\sigma_1, \sigma_2]$. From Barbara's decision problem in Table 9.1.3 it then follows that $\sigma_2(\textit{red}) > 0$. Thus, we see that $\sigma_2(\textit{blue}) > 0$ and $\sigma_2(\textit{red}) > 0$. This means that for Barbara, her choices *blue* and *red* must both be optimal under the same second-order expectation $e_2[\sigma_1, \sigma_2]$. In particular, her choices *blue* and *red* must yield the same expected utility under $e_2[\sigma_1, \sigma_2]$. These two expected utilities are given by

$$u_2(\textit{blue},\ e_2[\sigma_1, \sigma_2]) = \sigma_1(b) \cdot \sigma_2(r) \cdot 2 + \sigma_1(g) \cdot \sigma_2(b) \cdot 2 + \sigma_1(g) \cdot \sigma_2(r) \cdot 4$$

and

$$u_2(\textit{red},\ e_2[\sigma_1, \sigma_2]) = \sigma_1(b) \cdot \sigma_2(b) \cdot 6 + \sigma_1(b) \cdot \sigma_2(r) \cdot 3 + \sigma_1(g) \cdot \sigma_2(b) \cdot 6 + \sigma_1(g) \cdot \sigma_2(r) \cdot 3.$$

By setting the two expected utilities equal we get

$$\sigma_1(b) \cdot [6\sigma_2(b) + \sigma_2(r)] = \sigma_1(g) \cdot [-4\sigma_2(b) + \sigma_2(r)]. \tag{9.1.4}$$

Since $\sigma_1(b) > 0$, $\sigma_1(g) > 0$, $\sigma_2(b) > 0$ and $\sigma_2(r) > 0$, we conclude on the basis of (9.1.3) and (9.1.4) that

$$\begin{aligned} \sigma_1(g) \cdot [-4\sigma_2(b) + \sigma_2(r)] &= \sigma_1(b) \cdot [6\sigma_2(b) + \sigma_2(r)] \\ &> \sigma_1(b) \cdot [4\sigma_2(b) + \sigma_2(r)] = \sigma_1(g) \cdot [\sigma_2(b) + 4\sigma_2(r)]. \end{aligned}$$

This implies that $-4\sigma_2(b) + \sigma_2(r) > \sigma_2(b) + 4\sigma_2(r)$, which is impossible. Hence, we see that Case 1.1 leads to a contradiction, and therefore it cannot be that $\sigma_2(\textit{blue}) > 0$.

Case 1.2. Suppose that $\sigma_2(\textit{blue}) = 0$. This means that $\sigma_2(\textit{red}) = 1$. Substituting this into (9.1.3) yields $\sigma_1(b) = 4\sigma_1(g)$, and hence

$$\sigma_1(\textit{blue}) = 0.8 \text{ and } \sigma_1(\textit{green}) = 0.2.$$

It may be verified that the belief pair (σ_1, σ_2), where σ_1 is as above, and $\sigma_2(\textit{red}) = 1$, is a psychological Nash equilibrium.

Indeed, from (9.1.3) we know that your choices *blue* and *green* yield the same expected utility under $e_1[\sigma_1, \sigma_2]$, which is 3.6. As your choice *red* yields a lower expected utility under $e_1[\sigma_1, \sigma_2]$, we conclude that both of your choices *blue* and *green* are optimal under $e_1[\sigma_1, \sigma_2]$. It may also be verified that Barbara's choice *red* is optimal under the induced second-order expectation $e_2[\sigma_1, \sigma_2]$. We thus see that (σ_1, σ_2) is indeed a psychological Nash equilibrium.

Case 2. Suppose that $\sigma_1(\textit{blue}) = 0$. Then, it must be that $\sigma_1(\textit{green}) = 1$. Hence, your choice *green* must be optimal under the induced second-order expectation $e_1[\sigma_1, \sigma_2]$. From your decision problem in Table 9.1.3 we see that this is only possible if $\sigma_1(\textit{blue}) > 0$, which is a contradiction. Hence, it cannot be that $\sigma_1(\textit{blue}) = 0$.

Altogether, we conclude that there is only one psychological Nash equilibrium in this example, which is given by

$$\sigma_1(\textit{blue}) = 0.8,\ \sigma_1(\textit{green}) = 0.2 \text{ and } \sigma_2(\textit{red}) = 1.$$

In this psychological Nash equilibrium your optimal choices are *blue* and *green,* whereas Barbara's unique optimal choice is *red.* By Theorem 9.1.2 we thus conclude that with a simple belief hierarchy that expresses common belief in rationality, you can only rationally choose *blue* and *green,* whereas Barbara can only rationally choose *red.*

Moreover, you can only surprise Barbara with probability at most 0.8, by wearing *green,* whereas Barbara expects not to surprise you at all by wearing *red.* In contrast, if we would not insist on a simple belief hierarchy, then under common belief in rationality both you and Barbara can expect to surprise the other person with probability 1. This is seen most easily from the beliefs diagram in Figure 8.1.3, in Section 8.1.

Also here there is a clear intuition for this phenomenon: With a simple belief hierarchy, a player must believe that his opponent is correct about his actual first- and second-order belief. As his optimal choice is based on his first- and second order belief, the correct beliefs assumption severely restricts his possibilities of surprising the opponent. For Barbara it has rather drastic consequences: With the correct beliefs assumption she is no longer in a position to surprise you, not even with the slightest of probabilities.

9.1.4 Psychological Nash Equilibria Always Exist

So far we have been combining the notion of a simple belief hierarchy with the conditions of common belief in rationality. Similarly to Chapters 4 and 6, we ask the question whether this combination is always possible. That is, will there always be, for each of the players, a simple belief hierarchy that expresses common belief in rationality?

As we will see, the answer is "yes". In view of Theorem 9.1.1, the key lies in showing that a psychological Nash equilibrium always exists for every psychological game. This is the content of the following theorem.

Theorem 9.1.3 (Existence of psychological Nash equilibrium) *For every psychological game there is always at least one psychological Nash equilibrium.*

In Theorem 9.1.1 we have seen that every simple belief hierarchy that is generated by a psychological Nash equilibrium will express common belief in rationality. Since the theorem above guarantees that such a psychological Nash equilibrium will always exist, it follows that we can always find simple belief hierarchies that express common belief in rationality.

Theorem 9.1.4 (Simple belief hierarchies that express common belief in rationality) *For every psychological game, and every player i, there is always at least one simple belief hierarchy for player i that expresses common belief in rationality.*

In other words, combining the condition of correct beliefs with those of common belief in rationality never leads to logical contradictions in a psychological game.

9.2 Symmetric Beliefs

In this section we combine the conditions of common belief in rationality with the notion of a *symmetric* belief hierarchy. It turns out that the definition of a symmetric belief hierarchy, as defined in Section

4.2.1 for standard games, can also be used for psychological games. As a consequence, we inherit the result from Section 4.2.2 which states that the symmetric belief hierarchies are precisely those belief hierarchies that are induced by a *common prior* on choice-type combinations. We then introduce the concept of *psychological correlated equilibrium,* selecting those common priors that satisfy an optimality condition similar to that of *correlated equilibrium,* but now adapted to the class of psychological games. It is shown that a symmetric belief hierarchy expresses common belief in rationality precisely when it is induced by a psychological correlated equilibrium. In that sense, psychological correlated equilibrium is the counterpart to correlated equilibrium for the class of psychological games. We conclude by observing that every psychological Nash equilibrium induces a psychological correlated equilibrium. Since we know that a psychological Nash equilibrium always exists, the existence of a psychological correlated equilibrium is guaranteed as well.

9.2.1 Symmetric Belief Hierarchies and Common Prior

In Section 4.2.1 we have discussed, and formalized, the idea of a *symmetric belief hierarchy* for standard games. The intuitive idea is that, within a given beliefs diagram, there is a certain symmetry between the beliefs you have about the opponent, and the beliefs you believe the opponent to have about you. This idea, and its formalization, can be carried over *without any change* to psychological games. For completeness, we repeat the definition of a symmetric belief hierarchy here, adapted to the case of two players.

Definition 9.2.1 (Symmetric belief hierarchy) *(a) A **weighted beliefs diagram** starts from a beliefs diagram, removes the probabilities at the forked arrows (if there are any), and assigns to every arrow a from a choice c_i to an opponent's choice c_j some positive weight, which we call $w(a)$.*

*(b) Consider an arrow a from a choice c_i to an opponent's choice c_j. The **symmetric counterpart** to a is the arrow from the choice c_j to the choice c_i.*

*(c) A weighted beliefs diagram is **symmetric** if for every arrow a, the symmetric counterpart is also part of the diagram, and carries the same weight as a.*

(d) The weighted beliefs diagram induces a (normal) beliefs diagram in which the probability of an arrow a leaving a choice c_i is equal to

$$p(a) = \frac{w(a)}{\sum_{\text{arrows } a' \text{ leaving } c_i} w(a')}.$$

*(e) A belief hierarchy is **symmetric** if it is part of a beliefs diagram that is induced by a symmetric weighted beliefs diagram.*

In Section 4.2.2 it was shown that for standard games, the symmetric belief hierarchies are exactly those that are induced by a *common prior* on choice-type combinations. Since the belief hierarchies in psychological games are exactly the same as in two-player standard games, this result will hold for psychological games as well. For completeness, we repeat the definition of a common prior on choice-type combinations here, adapted to the case of two players.

Definition 9.2.2 (Common prior on choice-type combinations) *Consider a beliefs diagram in choice-type representation, with associated sets of types T_i for every player i. Let $C \times T$ be the*

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 9.2.1 Decision problems for "Dinner with a huge preference for surprise"

corresponding set of all choice-type combinations.

(a) A ***common prior on choice-type combinations*** *is a probability distribution π that assigns to every choice-type combination (c,t) in $C \times T$ a probability $\pi(c,t)$.*

(b) The beliefs diagram is ***induced by a common prior*** *π on choice-type combinations, if for every choice-type combination $((c_i,t_i),(c_j,t_j))$ and every player i, the corresponding arrow a from (c_i,t_i) to (c_j,t_j) is present exactly when $\pi((c_i,t_i),(c_j,t_j)) > 0$, and the probability of this arrow a is equal to*

$$p(a) = \frac{\pi((c_i,t_i),(c_j,t_j))}{\pi(c_i,t_i)}.$$

(c) A belief hierarchy is ***induced by a common prior*** *π on choice-type combinations if it is part of a beliefs diagram that is induced by π.*

Theorem 4.2.1 stated that in every standard game, the symmetric belief hierarchies are precisely those that are induced by a common prior on choice-type combinations. Since the definitions of a symmetric belief hierarchy and common prior remain exactly the same when we move to psychological games, Theorem 4.2.1 applies to psychological games as well.

9.2.2 Relation with Psychological Correlated Equilibrium

Consider a symmetric belief hierarchy induced by a common prior π on choice-type combinations. Suppose we impose, in addition, the conditions of common belief in rationality. What conditions does this impose on the common prior π? This is the question we wish to answer in this subsection.

Consider a symmetric belief hierarchy β_i for player i induced by the common prior π on choice-type combinations. Suppose that, within a beliefs diagram in choice-type representation, the belief hierarchy β_i starts at some choice-type pair (c_i^*, t_i^*). Assume, in addition, that β_i expresses common belief in rationality. Then, in particular, the belief hierarchy β_i believes in opponent j's rationality. That is, if β_i's first-order belief assigns a positive probability to an opponent's choice-type pair (c_j^*, t_j^*), then c_j^* must be optimal for j, given what i believes is j's second-order expectation conditional on (c_j^*, t_j^*).

But what is j's second-order expectation induced by the belief hierarchy β_i when we condition on (c_j^*, t_j^*) ? To illustrate this, let us go back to the example "Dinner with a huge preference for surprise" from Section 8.4.7. For convenience, we reproduce the decision problems for you and Barbara in Table 9.2.1. Consider the common prior π on choice-type combination given by

$$\pi((\textit{black}, t_1^b),(\textit{black}, t_2^b)) = 0.2,\ \pi((\textit{black}, t_1^b),(\textit{white}, t_2^w)) = 0.2,$$
$$\pi((\textit{white}, t_1^w),(\textit{black}, t_2^b)) = 0.4 \text{ and } \pi((\textit{white}, t_1^w),(\textit{white}, t_2^w)) = 0.2. \tag{9.2.1}$$

It may be verified that this common prior is induced by the symmetric weighted beliefs diagram in choice-type representation in the upper half of Figure 9.2.1, which in turn induces the symmetric beliefs diagram in choice-type representation in the lower half of that figure.

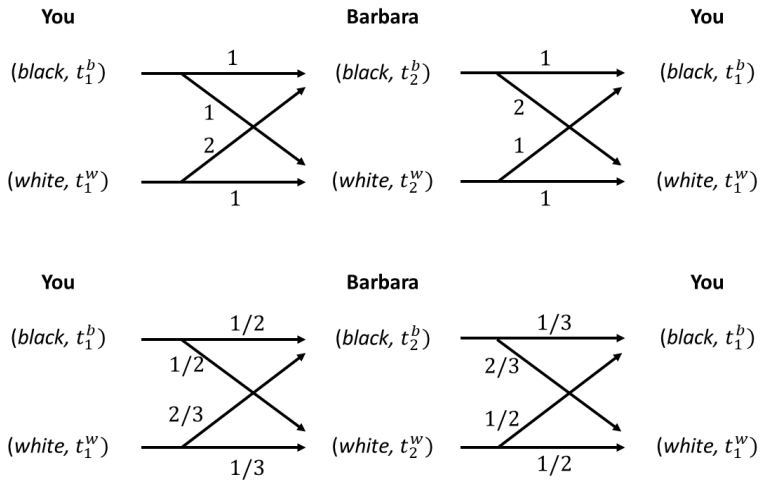

Figure 9.2.1 Symmetric beliefs diagram for "Dinner with a huge preference for surprise"

Consider your symmetric belief hierarchy that starts at your choice-type pair $(black,\ t_1^b)$. In your first-order belief, you assign probability 1/2 to Barbara's choice-type pair $(black,\ t_2^b)$. If you believe in Barbara's rationality, then Barbara's choice *black* must be optimal for her, given what you believe is Barbara's second-order expectation conditional on $(black,\ t_2^b)$. By following the arrows for two steps, starting at her choice-type pair $(black,\ t_2^b)$, we see that Barbara's second-order expectation on choice-type combinations conditional on her choice-type pair $(black,\ t_2^b)$ is given by $e_2(\cdot \mid \pi,\ (black_2,\ t_2^b))$ where

$$\begin{aligned}
e_2((black_1, t_1^b),\ (black_2, t_2^b) \mid \pi, (black_2, t_2^b)) &= 1/3 \cdot 1/2 = 1/6, \\
e_2((black_1, t_1^b),\ (white_2, t_2^w) \mid \pi, (black_2, t_2^b)) &= 1/3 \cdot 1/2 = 1/6, \\
e_2((white_1, t_1^w),\ (black_2, t_2^b) \mid \pi, (black_2, t_2^b)) &= 2/3 \cdot 2/3 = 4/9 \text{ and} \\
e_2((white_1, t_1^w),\ (white_2, t_2^w) \mid \pi, (black_2, t_2^b)) &= 2/3 \cdot 1/3 = 2/9. \qquad (9.2.2)
\end{aligned}$$

Here, the subindices 1 and 2 indicate whether the choice belongs to you (player 1) or to Barbara (player 2).

Concentrate, for the moment, on the conditional probability

$$e_2((white_1, t_1^w),\ (white_2, t_2^w) \mid \pi, (black_2, t_2^b)) = 2/3 \cdot 1/3 = 2/9.$$

By looking at the beliefs diagram, we see that

$$2/3 = \pi((white_1, t_1^w) \mid (black_2, t_2^b))$$

and that

$$1/3 = \pi((white_2, t_2^w) \mid (white_1, t_1^w)).$$

Thus,

$$e_2((white_1, t_1^w),\ (white_2, t_2^w) \mid \pi, (black_2, t_2^b)) = \pi((white_1, t_1^w) \mid (black_2, t_2^b)) \cdot \pi((white_2, t_2^w) \mid (white_1, t_1^w)).$$

This makes intuitive sense: Conditional on $(black_2, t_2^b)$, the probability that Barbara assigns to the event that your choice-type pair is $(white_1, t_1^w)$ and that you believe that her choice-type pair is

$(white_2, t_2^w)$ consists of the conditional probability that she assigns to your choice-type pair $(white_1, t_1^w)$, multiplied by the probability that you assign to Barbara's choice-type pair $(white_2, t_2^w)$ conditional on your choice-type pair being $(white_1, t_1^w)$.

This expression can be generalized: Consider a common prior π on choice-type combinations, a player i and some choice-type pair (c_i^*, t_i^*) for player i to which π assigns positive probability. Then, the induced second-order expectation (on choice-type combinations) for player i conditional on (c_i^*, t_i^*) is given by $e_i(\cdot \mid \pi, (c_i^*, t_i^*))$ where

$$e_i((c_j, t_j), (c_i, t_i) \mid \pi, (c_i^*, t_i^*)) := \pi((c_j, t_j) \mid (c_i^*, t_i^*)) \cdot \pi((c_i, t_i) \mid (c_j, t_j))$$

for every choice-type pair (c_j, t_j) for player j and every choice-type pair (c_i, t_i) for player i.

Now, consider again a symmetric belief hierarchy β_i for player i, induced by a common prior π on choice-type combinations. Suppose that in the first-order belief, β_i assigns positive probability to j's choice-type pair (c_j, t_j). If player i believes in j's rationality, then the choice c_j must be optimal for player j given the induced second-order expectation $e_j(\cdot \mid \pi, (c_j, t_j))$. Indeed, within the symmetric belief hierarchy β_i induced by π, player i believes that j chooses c_j because i believes that j holds the second-order expectation $e_j(\cdot \mid \pi, (c_j, t_j))$.

Suppose next that i also believes that j believes in i's rationality. Assume that in the second-order belief, i believes that j assigns a positive probability to (c_i, t_i). Then, by a similar argument as above, the choice c_i must be optimal for player i given the induced second-order expectation $e_i(\cdot \mid \pi, (c_i, t_i))$.

A common prior π on choice-type combinations with these properties is called a *psychological correlated equilibrium.*

Definition 9.2.3 (Psychological correlated equilibrium) *A common prior π on choice-type combinations is a **psychological correlated equilibrium** if for every player i, and every choice-type pair (c_i, t_i) with $\pi(c_i, t_i) > 0$, the choice c_i is optimal for the induced second-order expectation $e_i(\cdot \mid \pi, (c_i, t_i))$ of player i conditional on his choice-type pair (c_i, t_i).*

Note the similarity with the definition of a correlated equilibrium for standard games in Definition 4.2.3. The only difference is that in a correlated equilibrium for standard games, the choice c_i must be optimal for the induced (first-order) belief $\pi(\cdot \mid (c_i, t_i))$ about player j's choice-type pair, whereas in a psychological correlated equilibrium the choice c_i must be optimal for the induced second-order expectation $e_i(\cdot \mid \pi, (c_i, t_i))$ about player j's choice-type pair *and player i's choice-type pair.* The reason is that in a psychological game, player i's preferences depend on his second-order expectation, and not only on his first-order belief.

To illustrate the definition, let us return to the example "Dinner with a huge preference for surprise", and the common prior π on choice-type combinations as given by (9.2.1) above. We will show that π is a psychological correlated equilibrium.

Note that π assigns positive probability to the choice-type pairs $(black_1, t_1^b)$, $(white_1, t_1^w)$, $(black_2, t_2^b)$ and $(white_2, t_2^w)$. We first show that Barbara's choice $black_2$ is optimal for the induced second-order expectation $e_2(\cdot \mid \pi, (black_2, t_2^b))$. Recall that this second-order expectation is given by (9.2.2) above. The expected utilities for Barbara of choosing $black_2$ and $white_2$ are therefore given by

$$u_2(black_2,\ e_2(\cdot \mid \pi, (black_2, t_2^b)) = 2 \text{ and}$$
$$u_2(white_2,\ e_2(\cdot \mid \pi, (black_2, t_2^b)) = \tfrac{1}{6} \cdot 8 + \tfrac{1}{6} \cdot 0 + \tfrac{4}{9} \cdot 0 + \tfrac{2}{9} \cdot 0 = \tfrac{8}{6},$$

which implies that her choice $black_2$ is indeed optimal for the induced second-order expectation $e_2(\cdot \mid \pi, (black_2, t_2^b))$.

Question 9.2.1 *Show that Barbara's choice $white_2$ is optimal for the induced second-order expectation $e_2(\cdot \mid \pi, (white_2, t_2^w))$.*

In a similar fashion it can be shown that your choice $black_1$ is optimal for the second-order expectation $e_1(\cdot \mid \pi, (black_1, t_1^b))$ and that your choice $white_1$ is optimal for the second-order expectation $e_1(\cdot \mid \pi, (white_1, t_1^w))$. As such, we conclude that the common prior π is a psychological correlated equilibrium.

We have seen above that if a symmetric belief hierarchy is induced by a common prior π on choice-type combinations, and the belief hierarchy expresses common belief in rationality, then π must be a psychological correlated equilibrium. We will now show that the other direction is also true: If a symmetric belief hierarchy is induced by a psychological correlated equilibrium, then the belief hierarchy will always express common belief in rationality.

To see this, consider, within a beliefs diagram in choice-type representation, a symmetric belief hierarchy β_i for player i induced by a psychological correlated equilibrium π. We first show that β_i believes in j's rationality. Assume that β_i assigns a positive probability to j's choice-type pair (c_j, t_j). Then, in particular, $\pi(c_j, t_j) > 0$. Since π is a psychological correlated equilibrium, choice c_j must be optimal for the induced second-order expectation $e_j(\cdot \mid \pi, (c_j, t_j))$. By construction, player i believes in β_i that, conditional on j's choice-type pair (c_j, t_j), player j's second-order expectation is given by $e_j(\cdot \mid \pi, (c_j, t_j))$. Thus, if β_i assigns a positive probability to (c_j, t_j), then the choice c_j is optimal, given what player i believes about j's second-order expectation. As such, β_i believes in j's rationality.

We next show that β_i believes that j believes in i's rationality. Suppose that β_i assigns, in its second-order belief, a positive probability to the choice-type pair (c_i, t_i). Then, β_i's first-order belief assigns a positive probability to some choice-type pair (c_j, t_j), and the conditional belief $\pi(\cdot \mid (c_j, t_j))$ assigns a positive probability to (c_i, t_i). In particular, $\pi(c_i, t_i) > 0$. Since π is a psychological correlated equilibrium, the choice c_i is optimal for the induced second-order expectation $e_i(\cdot \mid (c_i, t_i))$. Now, player j believes that, conditional on (c_i, t_i), player i's second-order expectation is given by $e_i(\cdot \mid (c_i, t_i))$. Thus, player i believes that player j believes that the choice c_i is optimal, given what player i believes that player j believes about i's second-order expectation, conditional on (c_i, t_i)). In other words, player i believes that player j believes in i's rationality.

In a similar way, it can be shown that β_i also expresses 3-fold belief in rationality and higher, and thus that β_i expresses common belief in rationality. By putting all our insights together, we arrive at the following conclusion.

Theorem 9.2.1 (Relation with psychological correlated equilibrium) *A belief hierarchy is symmetric and expresses common belief in rationality, if and only if, the belief hierarchy is induced by a psychological correlated equilibrium.*

Eventually, we are interested in the *choices* you can rationally make if you hold a symmetric belief hierarchy that expresses common belief in rationality. Suppose you hold a symmetric belief hierarchy β_i that expresses common belief in rationality. Then, we know from Theorem 9.2.1 that the belief hierarchy β_i is induced by a psychological correlated equilibrium π. If the belief hierarchy starts at the choice-type pair (c_i, t_i), then your second-order expectation is given by $e_i(\cdot \mid \pi, (c_i, t_i))$. Hence, the choices that are optimal for you are precisely the choices that are optimal for the induced second-order expectation $e_i(\cdot \mid \pi, (c_i, t_i))$. Such choices are said to be *optimal in a psychological correlated equilibrium.*

Definition 9.2.4 (Choice optimal in a psychological correlated equilibrium) *A choice c_i^* is* ***optimal in a psychological correlated equilibrium*** *if there is a psychological correlated equi-*

librium π, and a choice-type pair (c_i, t_i) with $\pi(c_i, t_i) > 0$, such that the choice c_i^* is optimal for the induced second-order expectation $e_i(\cdot \mid \pi, (c_i, t_i))$.

By the argument above, we thus know that every choice that is optimal for a symmetric belief hierarchy that expresses common belief in rationality must be optimal in a psychological correlated equilibrium.

The converse is also true: Take a choice c_i^* that is optimal in a psychological correlated equilibrium. Then, there is a psychological correlated equilibrium π, and a choice-type pair (c_i, t_i) with $\pi(c_i, t_i) > 0$, such that c_i^* is optimal for the induced second-order expectation $e_i(\cdot \mid \pi, (c_i, t_i))$. Now, consider the belief hierarchy β_i induced by the psychological correlated equilibrium π that starts at (c_i, t_i). By Theorem 9.2.1 we know that β_i is symmetric and expresses common belief in rationality. Moreover, by construction, the second-order expectation induced by β_i is $e_i(\cdot \mid \pi, (c_i, t_i))$. Thus, the choice c_i^* is optimal for the belief hierarchy β_i that is symmetric and expresses common belief in rationality. These two insights lead to the following result.

Theorem 9.2.2 (Relation with psychological correlated equilibrium choices) *A choice is optimal for a symmetric belief hierarchy that expresses common belief in rationality, if and only if, the choice is optimal in a psychological correlated equilibrium.*

An important question is whether we can always find a symmetric belief hierarchy that expresses common belief in rationality. The answer is "yes", and the reason is simple: In Theorem 9.1.4 we have seen that there is always a *simple* belief hierarchy that expresses common belief in rationality. Moreover, we know from Theorem 4.3.5 that every simple belief hierarchy is symmetric. It therefore follows that we can always find a symmetric belief hierarchy that expresses common belief in rationality. By combining this insight with Theorem 9.2.1, we conclude moreover that in every psychological game there is at least one psychological correlated equilibrium.

Theorem 9.2.3 (Existence) *For every psychological game there is at least one psychological correlated equilibrium. Moreover, for every player there is at least one belief hierarchy that is symmetric and expresses common belief in rationality.*

That is, combining the condition of a symmetric belief hierarchy with the conditions of common belief in rationality never leads to logical contradictions.

9.2.3 Relation with Canonical Psychological Correlated Equilibrium

Recall from Section 4.3 that a belief hierarchy uses *one theory per choice* if it is generated by a beliefs diagram where every choice of a player only appears once. That is, for every choice there is exactly *one belief hierarchy,* or *theory,* that is used to support that choice. We have seen in Theorem 4.3.2 that a belief hierarchy is symmetric and uses one theory per choice, precisely when it is induced by a common prior on choice combinations. For convenience, we repeat here the definition of a common prior on choice combinations, and what it means for a belief hierarchy to be induced by a common prior on choice combinations. All is adapted to the case of two players.

Definition 9.2.5 (Common prior on choice combinations) *(a) A **common prior on choice combinations** is a probability distribution $\hat{\pi}$ that assigns to every choice pair (c_1, c_2) a probability $\hat{\pi}(c_1, c_2)$.*

(b) A beliefs diagram is ***induced by the common prior*** *on choice combinations $\hat{\pi}$ if every choice only appears once, if for every choice c_i and opponent's choice c_j, the arrow a from c_i to c_j is present exactly when $\hat{\pi}(c_i, c_j) > 0$, and this arrow a has probability*

$$p(a) = \frac{\hat{\pi}(c_i, c_j)}{\hat{\pi}(c_i)}.$$

(c) A belief hierarchy is ***induced by a common prior*** *on choice combinations $\hat{\pi}$ if it is part of a beliefs diagram induced by $\hat{\pi}$.*

For instance, all belief hierarchies in the beliefs diagram of Figure 9.2.1 are symmetric and use one theory per choice.

In general, consider a symmetric belief hierarchy β_i that uses one theory per choice. Then, we know from Theorem 4.3.2 that β_i is generated by a common prior on choice combinations. Suppose that, in addition, we impose that the belief hierarchy expresses common belief in rationality. What conditions does this put on the common prior on choice combinations?

From Theorem 9.2.1 we know that β_i is generated by a psychological correlated equilibrium π, which is a common prior on choice-type combinations. Moreover, as β_i uses one theory per choice, we know that for every choice c_i there is *exactly one* type $t_i^{c_i}$ such that $(c_i, t_i^{c_i})$ enters the beliefs diagram that generates β_i. But then, the psychological correlated equilibrium π can naturally be transformed into the common prior $\hat{\pi}$ on choice combinations where

$$\hat{\pi}(c_1, c_2) := \pi((c_1, t_1^{c_1}), (c_2, t_2^{c_2}))$$

for every choice pair (c_1, c_2).

Now, suppose that $\hat{\pi}(c_i^*) > 0$. Then, $\pi(c_i^*, t_i^{c_i^*}) > 0$. Since π is a psychological correlated equilibrium, c_i^* must be optimal for the second-order expectation $e_i(\cdot \mid \pi, (c_i^*, t_i^{c_i^*}))$ induced by π. Recall that this is a second-order expectation on choice-type combinations. This second-order expectation naturally induces the second-order expectation $e_i(\cdot \mid \hat{\pi}, c_i^*)$ on choice pairs (without types) generated by $\hat{\pi}$, given by

$$e_i((c_j, c_i) \mid \hat{\pi}, c_i) := \hat{\pi}(c_j \mid c_i^*) \cdot \hat{\pi}(c_i \mid c_j).$$

Then, stating that c_i^* is optimal for $e_i(\cdot \mid \pi, (c_i^*, t_i^{c_i^*}))$ is equivalent to saying that c_i^* is optimal for $e_i(\cdot \mid \hat{\pi}, c_i^*)$.

We thus see that the belief hierarchy β_i is induced by a common prior $\hat{\pi}$ on choice combinations, with the property that $\hat{\pi}(c_i) > 0$ only if the choice c_i is optimal for the induced second-order expectation $e_i(\cdot \mid \hat{\pi}, c_i)$. Such a common prior is called a *canonical psychological correlated equilibrium.*

Definition 9.2.6 (Canonical psychological correlated equilibrum) *A common prior $\hat{\pi}$ on choice combinations is a* ***canonical psychological correlated equilibrium*** *if for every player i, and every choice c_i with $\hat{\pi}(c_i) > 0$, the choice c_i is optimal for the induced second-order expectation $e_i(\cdot \mid \hat{\pi}, c_i)$ of player i conditional on his choice c_i.*

As an illustration, consider the example "Dinner with a huge preference for surprise", and the common prior $\hat{\pi}$ on choice combinations given by

$$\begin{aligned} &\hat{\pi}(\mathit{black}_1,\ \mathit{black}_2) = 0.2,\ \hat{\pi}(\mathit{black}_1,\ \mathit{white}_2) = 0.2, \\ &\hat{\pi}(\mathit{white}_1,\ \mathit{black}_2) = 0.4 \text{ and } \hat{\pi}(\mathit{white}_1,\ \mathit{white}_2) = 0.2. \end{aligned} \tag{9.2.3}$$

Note that this common prior $\hat{\pi}$ on choice combinations is induced by the common prior π on choice-type combinations as given in (9.2.1). As π only assigns positive probability to choice-type pairs where the same choice is never assigned to two different types, the common prior π induces symmetric belief hierarchies that use one theory per choice. In fact, it induces the same belief hierarchies as $\hat{\pi}$. We have seen above that π is a psychological correlated equilibrium. By using similar arguments, it can be shown that the common prior $\hat{\pi}$ on choice combinations is a *canonical* psychological correlated equilibrium.

Above we have argued that every symmetric belief hierarchy that uses one theory per choice and expresses common belief in rationality is induced by a canonical psychological correlated equilibrium. By using the same type of arguments as in Section 9.2.2, it can be shown that the converse is also true: Every belief hierarchy that is induced by a canonical psychological correlated equilibrium is symmetric, uses one theory per choice, and expresses common belief in rationality. By putting these two insights together we obtain the following characterization.

Theorem 9.2.4 (Relation with canonical psychological correlated equilibrium) *A belief hierarchy is symmetric, uses one theory per choice and expresses common belief in rationality, if and only if, the belief hierarchy is induced by a canonical psychological correlated equilibrium.*

As in Section 9.2.2, it is now easy to characterize the choices that can rationally be made with a symmetric belief hierarchy that uses one theory per choice and expresses common belief in rationality. These are precisely the choices that are optimal in a canonical psychological correlated equilibrium, which is defined as follows.

Definition 9.2.7 (Choice optimal in a canonical psychological correlated equilibrium) *A choice c_i^* is **optimal in a canonical psychological correlated equilibrium** if there is a canonical psychological correlated equilibrium $\hat{\pi}$, and a choice c_i with $\hat{\pi}(c_i) > 0$, such that the choice c_i^* is optimal for the induced second-order expectation $e_i(\cdot \mid \hat{\pi}, c_i)$.*

By similar arguments as those used in Section 9.2.2, we arrive at the following conclusion.

Theorem 9.2.5 (Relation with canonical psychological correlated equilibrium choices) *A choice is optimal for a symmetric belief hierarchy that uses one theory per choice and expresses common belief in rationality, if and only if, the choice is optimal in a canonical psychological correlated equilibrium.*

Finally, it can be shown that we can always find, for every player, a symmetric belief hierarchy that uses one theory per choice and expresses common belief in rationality. Indeed, in Theorem 9.1.4 we have seen that there is always a *simple* belief hierarchy that expresses common belief in rationality. Since we know from Theorem 4.3.5 that every simple belief hierarchy is symmetric and uses one theory per choice, it follows that we can always find a symmetric belief hierarchy that uses one theory per choice and expresses common belief in rationality. If we combine this result with Theorem 9.2.4, it follows that in every psychological game there is at least one canonical psychological correlated equilibrium.

Theorem 9.2.6 (Existence) *For every psychological game there is at least one canonical psychological correlated equilibrium. Moreover, for every player there is at least one belief hierarchy that is symmetric, uses one theory per choice and expresses common belief in rationality.*

As such, we can always combine the restrictions of a symmetric belief hierarchy, the one theory per choice condition, and the conditions of common belief in rationality without arriving at logical contradictions.

9.2.4 Examples

We will now identify, for each of the examples we have explored so far in this and the previous chapter, those choices you and Barbara can rationally make if you hold a symmetric, or simple, belief hierarchy that expresses common belief in rationality. For convenience, we will restrict attention to belief hierarchies that use one theory per choice. In the examples "The black and white dinner", "The black and white dinner with a twist" and "Dinner with a strong preference for surprise" the answer is clear: For each of these examples we have shown in Chapter 8 that under common belief in rationality you can only rationally choose to dress in *white* and that Barbara can only rationally choose to dress in *black*. This will still be the case if, in addition, we require you to have a symmetric, or simple, belief hierarchy. The same applies to the example "Exceeding Barbara's expectation" from Section 8.5.3, where we have shown that under common belief in rationality you and Barbara can only decide to practice for 1 week. Insisting on a symmetric, or simple, belief hierarchy will still single out this unique choice for you and Barbara.

Example 9.4: Dinner with a huge preference for surprise.

The decision problems for you and Barbara can be found in Table 9.2.1. Consider the beliefs diagram in the lower half of Figure 9.2.1. All belief hierarchies in this beliefs diagram are symmetric, use one theory per choice, and are induced by the canonical psychological correlated equilibrium $\hat{\pi}$ given by (9.2.3). By Theorem 9.2.4 we thus know that all these belief hierarchies express common belief in rationality. Since your choice *black* is optimal for the belief hierarchy that starts at $(black_1, t_1^b)$ and your choice *white* is optimal for the belief hierarchy that starts at $(white_1, t_1^w)$, it follows that under common belief in rationality with a symmetric belief hierarchy that uses one theory per choice, you can rationally choose *black* or *white*. Similarly for Barbara.

What if we insist on a simple belief hierarchy? Consider the pair of beliefs (σ_1, σ_2) where

$$\sigma_1 = (0.5) \cdot \ black \ + \ (0.5) \cdot \ white \text{ and } \sigma_2 = (0.5) \cdot \ black \ + \ (0.5) \cdot \ white.$$

Question 9.2.2 *Explain why (σ_1, σ_2) is a psychological Nash equilibrium.*

In the solution to this question, you have shown that both *black* and *white* are optimal for you in the psychological Nash equilibrium (σ_1, σ_2). Thus, by Theorem 9.1.2, you can rationally choose *black* and *white* under common belief in rationality with a simple belief hierarchy. Similarly for Barbara.

***Example 9.5: Barbara's birthday.**

Recall the decision problems for you and Barbara in Table 9.1.1. We have seen in Chapter 8 that under common belief in rationality you can only rationally buy a *necklace* or a *ring* and that Barbara can only rationally guess *necklace* or *ring*. Moreover, we have shown in Section 9.1.3 that you can also rationally buy these two items under common belief in rationality with a simple belief hierarchy. Since every simple belief hierarchy is symmetric and uses one theory per choice, we conclude that you can rationally buy a *necklace* or a *ring* under common belief in rationality with a symmetric belief hierarchy that uses one theory per choice.

For Barbara, however, the condition of imposing a simple belief hierarchy does rule out some of the choices she could rationally make under common belief in rationality. Indeed, in Chapter 8 we have seen that under common belief in rationality Barbara can rationally guess *necklace* or *ring*. In Section 9.1.3, however, we have shown that the unique psychological Nash equilibrium in this game is given by

$$\sigma_1 = (0.6) \cdot \ necklace + (0.4) \cdot ring \text{ and } \sigma_2 = \ necklace.$$

In this psychological Nash equilibrium, Barbara assigns probabilities 0.6 and 0.4 to you buying a *necklace* or a *ring,* respectively. Therefore, the only optimal choice for Barbara in a psychological Nash equilibrium is to guess that you buy a *necklace.* By Theorem 9.1.2 we thus know that under common belief in rationality with a simple belief hierarchy, she can only rationally guess *necklace.*

The question remains: Which choice(s) can Barbara rationally make under common belief in rationality with a *symmetric* belief hierarchy that uses one theory per choice? In view of Theorem 9.2.5 these are precisely the choices that are optimal for Barbara in a canonical psychological correlated equilibrium.

We will show that there is a unique canonical psychological correlated equilibrium $\hat{\pi}$, and that $\hat{\pi}$ corresponds to the Nash equiibrium above. To make the calculations more compact, we denote the choices *necklace, ring* and *bracelet* by n, r and b from now on.

Suppose that $\hat{\pi}$ is a canonical psychological correlated equilibrium. Since the choices b_1 and b_2 cannot rationally be made under common belief in rationality, we conclude that $\hat{\pi}(b_1) = 0$ and $\hat{\pi}(b_2) = 0$.

We first show that $\hat{\pi}(r_2) = 0$. Assume, on the contrary, that $\hat{\pi}(r_2) > 0$. Then, r_2 must be optimal for $e_2(\cdot \mid \hat{\pi}, r_2)$. As Barbara's preferences only depend on her first-order belief, this means that r_2 must be optimal for $\hat{\pi}(\cdot \mid r_2)$. This, in turn, implies that

$$\hat{\pi}(r_1|r_2) \geq 0.5 \text{ and hence } \hat{\pi}(r_1, r_2) > 0. \tag{9.2.4}$$

In particular, $\pi(r_1) > 0$, which means that r_1 must be optimal for $e_1(\cdot \mid \hat{\pi}, r_1)$. The expected utilities for you of choosing n_1 and r_1 under this second-order expectation are

$$\begin{array}{rcl} u_1(n_1,\ e_1(\cdot \mid \hat{\pi}, r_1)) & = & 3 \cdot [\hat{\pi}(n_2|r_1) \cdot \hat{\pi}(r_1|n_2) + \hat{\pi}(r_2|r_1) \cdot \hat{\pi}(r_1|r_2)] \text{ and} \\ u_1(r_1,\ e_1(\cdot \mid \hat{\pi}, r_1)) & = & 2 \cdot [\hat{\pi}(n_2|r_1) \cdot \hat{\pi}(n_1|n_2) + \hat{\pi}(r_2|r_1) \cdot \hat{\pi}(n_1|r_2)]. \end{array} \tag{9.2.5}$$

Since r_1 is optimal for $e_1(\cdot \mid \hat{\pi}, r_1)$ we must have that $u_1(r_1,\ e_1(\cdot \mid \hat{\pi}, r_1)) \geq u_1(n_1,\ e_1(\cdot \mid \hat{\pi}, r_1))$. Moreover, it may be verified that

$$\hat{\pi}(n_2|r_1) \cdot \hat{\pi}(r_1|n_2) + \hat{\pi}(r_2|r_1) \cdot \hat{\pi}(r_1|r_2) = 1 - \hat{\pi}(n_2|r_1) \cdot \hat{\pi}(n_1|n_2) - \hat{\pi}(r_2|r_1) \cdot \hat{\pi}(n_1|r_2). \tag{9.2.6}$$

By combining (9.2.5) and (9.2.6) and using the fact that $u_1(r_1, e_1(\cdot \mid \hat{\pi}, r_1)) \geq u_1(n_1, e_1(\cdot \mid \hat{\pi}, r_1))$, we arrive at the conclusion that

$$\hat{\pi}(n_2|r_1) \cdot \hat{\pi}(n_1|n_2) + \hat{\pi}(r_2|r_1) \cdot \hat{\pi}(n_1|r_2) \geq 0.6. \tag{9.2.7}$$

We know by (9.2.4) that $\hat{\pi}(n_1|r_2) \leq 0.5$. Moreover, by (9.2.4), it holds that $\hat{\pi}(r_1, r_2) > 0$ and hence $\hat{\pi}(r_2|r_1) > 0$. As $\hat{\pi}(n_2|r_1) + \hat{\pi}(r_2|r_1) = 1$, inequality (9.2.7) can only hold if

$$\hat{\pi}(n_2|r_1) > 0 \text{ and } \hat{\pi}(n_1|n_2) > 0.6. \tag{9.2.8}$$

This implies that

$$\hat{\pi}(r_1, n_2) > 0 \text{ and } \hat{\pi}(n_1, n_2) > 0. \tag{9.2.9}$$

In particular, $\hat{\pi}(n_1) > 0$. As such, n_1 must be optimal for $e_1(\cdot \mid \hat{\pi}, n_1)$. The expected utilities of choosing n_1 and r_1 under $e_1(\cdot \mid \hat{\pi}, n_1)$ are given by

$$\begin{array}{rcl} u_1(n_1,\ e_1(\cdot \mid \hat{\pi}, n_1)) & = & 3 \cdot [\hat{\pi}(n_2|n_1) \cdot \hat{\pi}(r_1|n_2) + \hat{\pi}(r_2|n_1) \cdot \hat{\pi}(r_1|r_2)] \text{ and} \\ u_1(r_1,\ e_1(\cdot \mid \hat{\pi}, n_1)) & = & 2 \cdot [\hat{\pi}(n_2|n_1) \cdot \hat{\pi}(n_1|n_2) + \hat{\pi}(r_2|n_1) \cdot \hat{\pi}(n_1|r_2)]. \end{array} \tag{9.2.10}$$

Since n_1 is optimal for $e_1(\cdot \mid \hat{\pi}, n_1)$ we must have that $u_1(n_1,\, e_1(\cdot \mid \hat{\pi}, n_1)) \geq u_1(r_1,\, e_1(\cdot \mid \hat{\pi}, n_1))$. Moreover, it may be verified that

$$\hat{\pi}(n_2|n_1) \cdot \hat{\pi}(n_1|n_2) + \hat{\pi}(r_2|n_1) \cdot \hat{\pi}(n_1|r_2) = 1 - \hat{\pi}(n_2|n_1) \cdot \hat{\pi}(r_1|n_2) - \hat{\pi}(r_2|n_1) \cdot \hat{\pi}(r_1|r_2). \quad (9.2.11)$$

By combining (9.2.10) and (9.2.11) and using the fact that $u_1(n_1,\, e_1(\cdot \mid \hat{\pi}, n_1)) \geq u_1(r_1,\, e_1(\cdot \mid \hat{\pi}, n_1))$, we arrive at the conclusion that

$$\hat{\pi}(n_2|n_1) \cdot \hat{\pi}(r_1|n_2) + \hat{\pi}(r_2|n_1) \cdot \hat{\pi}(r_1|r_2) \geq 0.4. \quad (9.2.12)$$

By (9.2.8) we know that $\hat{\pi}(n_1|n_2) > 0.6$ and hence $\hat{\pi}(r_1|n_2) < 0.4$. As $\hat{\pi}(n_2|n_1) + \hat{\pi}(r_2|n_1) = 1$, it follows that (9.2.12) can only hold if $\hat{\pi}(r_2|n_1) > 0$, and hence

$$\hat{\pi}(n_1, r_2) > 0. \quad (9.2.13)$$

Let us go back to (9.2.7). If we multiply both sides with $\hat{\pi}(r_1) \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2)$, and use the fact that

$$\begin{array}{rcl} \hat{\pi}(n_2|r_1) \cdot \hat{\pi}(r_1) & = & \hat{\pi}(r_1, n_2),\ \hat{\pi}(n_1|n_2) \cdot \hat{\pi}(n_2) = \hat{\pi}(n_1, n_2), \\ \hat{\pi}(r_2|r_1) \cdot \hat{\pi}(r_1) & = & \hat{\pi}(r_1, r_2) \text{ and } \hat{\pi}(n_1|r_2) \cdot \hat{\pi}(r_2) = \hat{\pi}(n_1, r_2), \end{array}$$

it follows that

$$\begin{array}{rl} & \hat{\pi}(r_1, n_2) \cdot \hat{\pi}(n_1, n_2) \cdot \hat{\pi}(r_2) + \hat{\pi}(r_1, r_2) \cdot \hat{\pi}(n_1, r_2) \cdot \hat{\pi}(n_2) \\ \geq & 0.6 \cdot \hat{\pi}(r_1) \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2) \\ = & 0.6 \cdot [\hat{\pi}(r_1, n_2) + \hat{\pi}(r_1, r_2)] \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2) \\ = & 0.6 \cdot \hat{\pi}(r_1, n_2) \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2) + 0.6 \cdot \hat{\pi}(r_1, r_2) \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2). \end{array}$$

This can be simplied to

$$\hat{\pi}(r_1, n_2) \cdot \hat{\pi}(r_2) \cdot [\hat{\pi}(n_1, n_2) - 0.6 \cdot \hat{\pi}(n_2)] + \hat{\pi}(r_1, r_2) \cdot \hat{\pi}(n_2) \cdot [\hat{\pi}(n_1, r_2) - 0.6 \cdot \hat{\pi}(r_2)] \geq 0. \quad (9.2.14)$$

By definition,

$$\begin{array}{rcl} \hat{\pi}(n_1, n_2) - 0.6 \cdot \hat{\pi}(n_2) & = & \hat{\pi}(n_1, n_2) - 0.6 \cdot [\hat{\pi}(n_1, n_2) + \hat{\pi}(r_1, n_2)] \\ & = & 0.4 \cdot \hat{\pi}(n_1, n_2) - 0.6 \cdot \hat{\pi}(r_1, n_2), \end{array}$$

and

$$\begin{array}{rcl} \hat{\pi}(n_1, r_2) - 0.6 \cdot \hat{\pi}(r_2) & = & \hat{\pi}(n_1, r_2) - 0.6 \cdot [\hat{\pi}(n_1, r_2) + \hat{\pi}(r_1, r_2)] \\ & = & 0.4 \cdot \hat{\pi}(n_1, r_2) - 0.6 \cdot \hat{\pi}(r_1, r_2). \end{array}$$

Substituting this into (9.2.14) yields

$$\begin{array}{c} \hat{\pi}(r_1, n_2) \cdot \hat{\pi}(r_2) \cdot [0.4 \cdot \hat{\pi}(n_1, n_2) - 0.6 \cdot \hat{\pi}(r_1, n_2)] + \\ + \hat{\pi}(r_1, r_2) \cdot \hat{\pi}(n_2) \cdot [0.4 \cdot \hat{\pi}(n_1, r_2) - 0.6 \cdot \hat{\pi}(r_1, r_2)] \geq 0. \end{array} \quad (9.2.15)$$

Recall from (9.2.8) that $\hat{\pi}(n_1|n_2) > 0.6$. By definition,

$$\hat{\pi}(n_1|n_2) = \frac{\hat{\pi}(n_1, n_2)}{\hat{\pi}(n_2)} = \frac{\hat{\pi}(n_1, n_2)}{\hat{\pi}(n_1, n_2) + \hat{\pi}(r_1, n_2)} > 0.6$$

which implies that

$$\hat{\pi}(n_1, n_2) > 0.6 \cdot [\hat{\pi}(n_1, n_2) + \hat{\pi}(r_1, n_2)]$$

and hence

$$0.4 \cdot \hat{\pi}(n_1, n_2) - 0.6 \cdot \hat{\pi}(r_1, n_2) > 0. \tag{9.2.16}$$

Moreover, we know from (9.2.4) that $\hat{\pi}(r_1|r_2) \geq 0.5$. By definition,

$$\hat{\pi}(r_1|r_2) = \frac{\hat{\pi}(r_1, r_2)}{\hat{\pi}(r_2)} = \frac{\hat{\pi}(r_1, r_2)}{\hat{\pi}(r_1, r_2) + \hat{\pi}(n_1, r_2)} \geq 0.5$$

which implies that

$$\hat{\pi}(r_1, r_2) \geq 0.5 \cdot [\hat{\pi}(r_1, r_2) + \hat{\pi}(n_1, r_2)]$$

and hence

$$0.5 \cdot \hat{\pi}(n_1, r_2) - 0.5 \cdot \hat{\pi}(r_1, r_2) \leq 0.$$

Since we have seen in (9.2.4) that $\hat{\pi}(r_1, r_2) > 0$, it follows from the inequality above that

$$0.4 \cdot \hat{\pi}(n_1, r_2) - 0.6 \cdot \hat{\pi}(r_1, r_2) < 0. \tag{9.2.17}$$

In view of (9.2.16) and (9.2.17), inequality (9.2.15) can be reformulated as

$$\frac{\hat{\pi}(r_1, n_2)}{\hat{\pi}(r_1, r_2)} \geq \frac{\hat{\pi}(n_2) \cdot [0.6 \cdot \hat{\pi}(r_1, r_2) - 0.4 \cdot \hat{\pi}(n_1, r_2)]}{\hat{\pi}(r_2) \cdot [0.4 \cdot \hat{\pi}(n_1, n_2) - 0.6 \cdot \hat{\pi}(r_1, n_2)]}. \tag{9.2.18}$$

Note that $\hat{\pi}(r_1, r_2) > 0$ by (9.2.4), and hence both fractions are well-defined.

Let us now return to (9.2.12). If we multiply both sides with $\hat{\pi}(n_1) \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2)$, and use the fact that

$$\begin{aligned}
\hat{\pi}(n_2|n_1) \cdot \hat{\pi}(n_1) &= \hat{\pi}(n_1, n_2),\ \hat{\pi}(r_1|n_2) \cdot \hat{\pi}(n_2) = \hat{\pi}(r_1, n_2), \\
\hat{\pi}(r_2|n_1) \cdot \hat{\pi}(n_1) &= \hat{\pi}(n_1, r_2) \text{ and } \hat{\pi}(r_1|r_2) \cdot \hat{\pi}(r_2) = \hat{\pi}(r_1, r_2),
\end{aligned}$$

it follows that

$$\begin{aligned}
& \hat{\pi}(n_1, n_2) \cdot \hat{\pi}(r_1, n_2) \cdot \hat{\pi}(r_2) + \hat{\pi}(n_1, r_2) \cdot \hat{\pi}(r_1, r_2) \cdot \hat{\pi}(n_2) \\
\geq\ & 0.4 \cdot \hat{\pi}(n_1) \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2) \\
=\ & 0.4 \cdot [\hat{\pi}(n_1, n_2) + \hat{\pi}(n_1, r_2)] \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2) \\
=\ & 0.4 \cdot \hat{\pi}(n_1, n_2) \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2) + 0.4 \cdot \hat{\pi}(n_1, r_2) \cdot \hat{\pi}(n_2) \cdot \hat{\pi}(r_2).
\end{aligned}$$

This can be simplied to

$$\hat{\pi}(n_1, n_2) \cdot \hat{\pi}(r_2) \cdot [\hat{\pi}(r_1, n_2) - 0.4 \cdot \hat{\pi}(n_2)] + \hat{\pi}(n_1, r_2) \cdot \hat{\pi}(n_2) \cdot [\hat{\pi}(r_1, r_2) - 0.4 \cdot \hat{\pi}(r_2)] \geq 0. \tag{9.2.19}$$

By definition,

$$\begin{aligned}
\hat{\pi}(r_1, n_2) - 0.4 \cdot \hat{\pi}(n_2) &= \hat{\pi}(r_1, n_2) - 0.4 \cdot [\hat{\pi}(n_1, n_2) + \hat{\pi}(r_1, n_2)] \\
&= 0.6 \cdot \hat{\pi}(r_1, n_2) - 0.4 \cdot \hat{\pi}(n_1, n_2),
\end{aligned}$$

and

$$\begin{aligned}
\hat{\pi}(r_1, r_2) - 0.4 \cdot \hat{\pi}(r_2) &= \hat{\pi}(r_1, r_2) - 0.4 \cdot [\hat{\pi}(n_1, r_2) + \hat{\pi}(r_1, r_2)] \\
&= 0.6 \cdot \hat{\pi}(r_1, r_2) - 0.4 \cdot \hat{\pi}(n_1, r_2).
\end{aligned}$$

Substituting this into (9.2.19) yields

$$\begin{aligned}&\hat{\pi}(n_1,n_2)\cdot\hat{\pi}(r_2)\cdot[0.6\cdot\hat{\pi}(r_1,n_2)-0.4\cdot\hat{\pi}(n_1,n_2)]+\\&+\hat{\pi}(n_1,r_2)\cdot\hat{\pi}(n_2)\cdot[0.6\cdot\hat{\pi}(r_1,r_2)-0.4\cdot\hat{\pi}(n_1,r_2)]\geq 0.\end{aligned}\tag{9.2.20}$$

Recall from (9.2.16) and (9.2.17) that

$$0.6\cdot\hat{\pi}(r_1,n_2)-0.4\cdot\hat{\pi}(n_1,n_2)<0 \text{ and } 0.6\cdot\hat{\pi}(r_1,r_2)-0.4\cdot\hat{\pi}(n_1,r_2)>0.$$

Hence, inequality (9.2.20) can be reformulated as

$$\frac{\hat{\pi}(n_1,n_2)}{\hat{\pi}(n_1,r_2)}\leq\frac{\hat{\pi}(n_2)\cdot[0.6\cdot\hat{\pi}(r_1,r_2)-0.4\cdot\hat{\pi}(n_1,r_2)]}{\hat{\pi}(r_2)\cdot[0.4\cdot\hat{\pi}(n_1,n_2)-0.6\cdot\hat{\pi}(r_1,n_2)]}.\tag{9.2.21}$$

Note that $\hat{\pi}(n_1,r_2)>0$ by (9.2.13), and hence the two fractions are well-defined.

By combining (9.2.18) and (9.2.21) we conclude that

$$\frac{\hat{\pi}(r_1,n_2)}{\hat{\pi}(r_1,r_2)}\geq\frac{\hat{\pi}(n_1,n_2)}{\hat{\pi}(n_1,r_2)}$$

and hence

$$\frac{\hat{\pi}(r_1,n_2)}{\hat{\pi}(n_1,n_2)}\geq\frac{\hat{\pi}(r_1,r_2)}{\hat{\pi}(n_1,r_2)}.\tag{9.2.22}$$

Recall from (9.2.4) and (9.2.8) that $\hat{\pi}(r_1|r_2)\geq 0.5$ and $\hat{\pi}(n_1|n_2)>0.6$. Since

$$\frac{\hat{\pi}(r_1,n_2)}{\hat{\pi}(n_1,n_2)}=\frac{\hat{\pi}(r_1,n_2)/\hat{\pi}(n_2)}{\hat{\pi}(n_1,n_2)/\hat{\pi}(n_2)}=\frac{\hat{\pi}(r_1|n_2)}{\hat{\pi}(n_1|n_2)}$$

and

$$\frac{\hat{\pi}(r_1,r_2)}{\hat{\pi}(n_1,r_2)}=\frac{\hat{\pi}(r_1,r_2)/\hat{\pi}(r_2)}{\hat{\pi}(n_1,r_2)/\hat{\pi}(r_2)}=\frac{\hat{\pi}(r_1|r_2)}{\hat{\pi}(n_1|r_2)}$$

this implies that

$$\frac{\hat{\pi}(r_1,n_2)}{\hat{\pi}(n_1,n_2)}<1 \text{ and } \frac{\hat{\pi}(r_1,r_2)}{\hat{\pi}(n_1,r_2)}\geq 1.$$

This, however, contradicts (9.2.22). We thus conclude that $\hat{\pi}(r_2)>0$ is impossible in a canonical psychological correlated equilibrium.

We therefore conclude that $\hat{\pi}(r_2)=0$, and hence $\hat{\pi}(n_2)=1$. Then, n_2 must be optimal for Barbara under $\hat{\pi}(\cdot\mid n_2)$, which implies that $\hat{\pi}(n_1|n_2)\geq 0.5$. In particular, $\hat{\pi}(n_1)>0$, and hence n_1 must be optimal for you under $e_1(\cdot\mid\hat{\pi},n_1)$. As $\hat{\pi}(r_2)=0$, we must have that $\hat{\pi}(n_2|n_1)=1$, and hence $e_1(\cdot\mid\hat{\pi},n_1)$ is given by

$$e_1((n_2,n_1)|\hat{\pi},n_1)=\hat{\pi}(n_1|n_2) \text{ and } e_1((n_2,r_1)|\hat{\pi},n_1)=\hat{\pi}(r_1|n_2),$$

whereas $e_1(\cdot\mid\hat{\pi},n_1)$ assigns probability zero to the other two choice pairs. Thus, the expected utilities of choosing n_1 and r_1 under this second-order expectation are

$$\begin{aligned}u_1(n_1,\ e_1(\cdot|\hat{\pi},n_1)) &= \hat{\pi}(r_1|n_2)\cdot 3 \text{ and}\\ u_1(r_1,\ e_1(\cdot|\hat{\pi},n_1)) &= \hat{\pi}(n_1|n_2)\cdot 2.\end{aligned}\tag{9.2.23}$$

As n_1 must be optimal for you under $e_1(\cdot \mid \hat{\pi}, n_1)$, we must have that $u_1(n_1,\, e_1(\cdot \mid \hat{\pi}, n_1)) \geq u_1(r_1,\, e_1(\cdot \mid \hat{\pi}, n_1))$, which is only possible if $\hat{\pi}(r_1|n_2) > 0$. In particular, $\hat{\pi}(r_1) > 0$, which means that r_1 must be optimal for you under $e_1(\cdot \mid \hat{\pi}, r_1)$. As $\hat{\pi}(r_2) = 0$, we must have that $\hat{\pi}(n_2|r_1) = 1$, and hence $e_1(\cdot \mid \hat{\pi}, r_1)$ is given by

$$e_1((n_2, n_1)|\hat{\pi}, r_1) = \hat{\pi}(n_1|n_2) \text{ and } e_1((n_2, r_1)|\hat{\pi}, r_1) = \hat{\pi}(r_1|n_2),$$

whereas $e_1(\cdot \mid \hat{\pi}, r_1)$ assigns probability zero to the other two choice pairs. This means, however, that $e_1(\cdot \mid \hat{\pi}, r_1) = e_1(\cdot \mid \hat{\pi}, n_1)$. Since r_1 must be optimal for you under $e_1(\cdot \mid \hat{\pi}, r_1)$, we conclude that $u_1(r_1,\, e_1(\cdot \mid \hat{\pi}, n_1)) \geq u_1(n_1,\, e_1(\cdot \mid \hat{\pi}, n_1))$.

Since we have seen above that $u_1(n_1,\, e_1(\cdot \mid \hat{\pi}, n_1)) \geq u_1(r_1,\, e_1(\cdot \mid \hat{\pi}, n_1))$, it must be that $u_1(n_1,\, e_1(\cdot \mid \hat{\pi}, n_1)) = u_1(r_1,\, e_1(\cdot \mid \hat{\pi}, n_1))$. In view of (9.2.23) this can only be if

$$\hat{\pi}(r_1|n_2) \cdot 3 = \hat{\pi}(n_1|n_2) \cdot 2.$$

Since $\hat{\pi}(r_1|n_2) = 1 - \hat{\pi}(n_1|n_2)$, this yields $\hat{\pi}(n_1|n_2) = 0.6$, and hence $\hat{\pi}(r_1|n_2) = 0.4$. As $\hat{\pi}(r_2) = 0$, we conclude that $\hat{\pi}$ must be given by

$$\hat{\pi}(n_1, n_2) = 0.6 \text{ and } \hat{\pi}(r_1, n_2) = 0.4. \tag{9.2.24}$$

It may indeed be verified that this is a canonical psychological correlated equilibrium.

Hence, the only canonical psychological correlated equilibrium is given by (9.2.24). Note that it corresponds precisely to the unique psychological Nash equilibrium in this game. Since Barbara's unique first-order belief assigns probability 0.6 to you buying a *necklace,* Barbara can only rationally guess *necklace* with a symmetric belief hierarchy that uses one theory per choice and expresses common belief in rationality.

Example 9.6: Surprising Barbara.

Recall the decision problems for you and Barbara in Table 9.1.2. We have seen in Chapter 8 that under common belief in rationality you can only rationally wear the colors *blue* and *green.* In addition, we explained in Section 9.1.3 why under common belief in rationality with a simple belief hierarchy, you can still rationally wear these two colors. Since every simple belief hierarchy is symmetric and uses one theory per choice, we conclude that under common belief in rationality with a symmetric belief hierarchy that uses one theory per choice you can also rationally wear *blue* and *green.*

Let us now turn to Barbara. In Chapter 8 we saw that Barbara can rationally wear the colors *blue* and *red* under common belief in rationality. In Section 9.1.3 we argued, however, that under common belief in rationality with a simple belief hierarchy, Barbara can only rationally wear *red.* Now, what color(s) can she rationally wear under common belief in rationality with a *symmetric* belief hierarchy that uses one theory per choice?

To answer this question, consider the common prior $\hat{\pi}$ on choice combinations given by

$$\hat{\pi}(blue_1,\ blue_2) = 1/130,\ \hat{\pi}(blue_1,\ red_2) = 100/130,$$
$$\hat{\pi}(green_1,\ blue_2) = 4/130 \text{ and } \hat{\pi}(green_1,\ red_2) = 25/130.$$

Question 9.2.3 *Show that $\hat{\pi}$ is a canonical psychological correlated equilibrium.*

As $\hat{\pi}(blue_2) > 0$ and $\hat{\pi}(red_2) > 0$, it follows by the definition of a canonical psychological correlated equilibrium that Barbara's choice $blue_2$ is optimal for the induced second-order expectation $e_2(\cdot \mid \hat{\pi},$

Common belief in rationality with ...	**Optimal choices are those that ...**
...	survive iterated elimination of choices and second-order expectations
symmetric belief hierarchy	are optimal in a psychological correlated equilibrium
symmetric belief hierarchy using one theory per choice	are optimal in a canonical psychological correlated equilibrium
simple belief hierarchy	are optimal in a psychological Nash equilibrium

Table 9.3.1 Comparison of the concepts in Chapters 8 and 9

blue$_2$), and that Barbara's choice *red*$_2$ is optimal for the induced second-order expectation $e_2(\cdot \mid \hat{\pi}$, *red*$_2$). Hence, Barbara's choices *blue* and *red* are both optimal in a canonical psychological correlated equilibrium. By Theorem 9.2.5 we thus conclude that Barbara can rationally wear *blue* and *red* under common belief in rationality with a symmetric belief hierarchy that uses one theory per choice.

In particular, if we assume common belief in rationality, and move from imposing a symmetric belief hierarchy that uses one theory per choice to imposing a simple belief hierarchy, then Barbara can no longer rationally choose *blue.*

9.3 Comparison of the Concepts

Like we did in Chapters 4 and 6, we will compare the various concepts we have discussed in this and the previous chapter. In Chapter 8 we discussed the conditions of *common belief in rationality,* whereas in Chapter 9 we supplemented these with the condition of a simple belief hierarchy, a symmetric belief hierarchy and a symmetric belief hierarchy that uses one theory per choice. Table 9.3.1 characterizes, for each of these sets of conditions on the belief hierarchy, the choices that can rationally be made by the players. By comparing this table with Table 4.4.1 in Chapter 4 we see that the iterated elimination of choices and second-order expectations, psychological correlated equilibrium, canonical psychological correlated equilibrium and psychological Nash equilibrium are the *psychological games counterparts* to the iterated elimination of strictly dominated choices, correlated equilibrium, canonical correlated equilibrium and Nash equilibrium, respectively. Indeed, the conditions on the belief hierarchies underlying these concepts are essentially the same. The only difference lies in what it means for a choice to be optimal: In standard games the optimality of a choice is defined relative to the first-order belief, whereas in psychological games the optimality is defined relative to the second-order expectation. Apart from this difference, the conditions on the belief hierarchies are really the same.

We next compare which choices these different concepts induce for each of the examples we have explored in this and the previous chapter. We focus both on the choices for you and the choices for Barbara. An overview can be found in Table 9.3.2. The section in the first column indicates the section where the example has been introduced. In the other columns, Y means "you" and B means "Barbara". The sections in these columns indicate the section where it has been shown that these specific choices can rationally be made by you and Barbara under these conditions on the belief hierarchy. For some rows and columns these sections have not been specified, because the result follows from other columns. For instance, if it has been shown that common belief in rationality

	Choices you can rationally make under common belief in rationality with ...			
Example	...	a symmetric belief hierarchy	a symmetric belief hierarchy using one theory per choice	a simple belief hierarchy
Surprising Barbara (Section 8.1)	Y: blue, green B: blue, red (Section 8.1)	Y: blue, green B: blue, red	Y: blue, green B: blue, red (Section 9.2.4)	Y: blue, green B: red (Section 9.1.3)
The black and white dinner (Section 8.4.2)	Y: white B: black (Section 8.4.2)	Y: white B: black	Y: white B: black	Y: white B: black
The black and white dinner with a twist (Section 8.4.3)	Y: white B: black (Section 8.4.3)	Y: white B: black	Y: white B: black	Y: white B: black
Dinner with a strong preference for surprise (Section 8.4.7)	Y: white B: black (Section 8.4.7)	Y: white B: black	Y: white B: black	Y: white B: black
Dinner with a huge preference for surprise (Section 8.4.7)	Y: white, black B: white, black (Section 8.4.7)	Y: white, black B: white, black	Y: white, black B: white, black	Y: white, black B: white, black (Section 9.2.4)
Exceeding Barbara's expectation (Section 8.5.3)	Y: 1 B: 1 (Section 8.5.3)	Y: 1 B: 1	Y: 1 B: 1	Y: 1 B: 1
Barbara's birthday (Section 8.6.2)	Y: neckl, ring B: neckl, ring (Section 8.6.2)	Y: neckl, ring B: ?	Y: neckl, ring B: neckl (Section 9.2.4)	Y: neckl, ring B: neckl (Section 9.1.3)

Table 9.3.2 The four concepts in the various examples

already yields a unique choice for you and Barbara, then it follows that the same applies for the other columns as well. Or, if it has been shown that common belief in rationality yields the same choices as common belief in rationality with a simple belief hierarchy, then these same choices will also result if we impose common belief in rationality with a symmetric belief hierarchy (with or without one theory per choice). And so on.

For the example "Barbara's birthday", we do not know what choices Barbara can rationally make under common belief in rationality with a symmetric belief hierarchy (without insisting on one theory per choice). Under these conditions, it is possible that Barbara can only rationally guess that you buy a *necklace,* or that she can rationally guess that you buy a *necklace* or a *ring.* We do not know at this moment. In the table we have left out the example "Your birthday" from Section 8.6.2, because it is essentially the same as the example "Barbara's birthday" after exchanging the roles of Barbara and you.

9.4 Proofs

9.4.1 Proofs of Section 9.1

Proof of Theorem 9.1.1. See the arguments in Section 9.1.2. ■

Proof of Theorem 9.1.2. See the arguments in Section 9.1.2. ■

Proof of Theorem 9.1.3. Like in Chapter 4, we use Kakutani's fixed point theorem (Theorem 4.5.1) to prove this theorem. For both players i, let $\Delta(C_i)$ denote the set of probability distributions on C_i. So, every pair of beliefs (σ_1, σ_2) belongs to the set $\Delta(C_1) \times \Delta(C_2)$. By

$$A := \Delta(C_1) \times \Delta(C_2)$$

we denote the set of all such belief pairs. Hence, A is a subset of some linear space $\mathbf{R}^X$. Moreover, it may easily be verified that the set A is nonempty, compact and convex.

For every $(\sigma_1, \sigma_2) \in A$ and both players i, let $C_i^{opt}(\sigma_1, \sigma_2)$ be the set of choices $c_i \in C_i$ that are optimal for player i under the induced second-order expectation $e_i[\sigma_1, \sigma_2]$. By $\Delta(C_i^{opt}(\sigma_1, \sigma_2))$ we denote the set of probability distributions in $\Delta(C_i)$ that only assign positive probability to choices in $C_i^{opt}(\sigma_1, \sigma_2)$. Define now the correspondence C^{opt} from A to A, which assigns to every belief pair $(\sigma_1, \sigma_2) \in A$ the set of belief pairs

$$C^{opt}(\sigma_1, \sigma_2) := \Delta(C_1^{opt}(\sigma_1, \sigma_2)) \times \Delta(C_2^{opt}(\sigma_1, \sigma_2)),$$

which is a subset of $\Delta(C_1) \times \Delta(C_2)$, and hence is a subset of A.

It may easily be verified that the set $C^{opt}(\sigma_1, \sigma_2)$ is nonempty and convex for every (σ_1, σ_2). It thus follows that the correspondence C^{opt} is convex-valued. We now show that the correspondence C^{opt} is upper-semicontinuous. That is, we must show that for every sequence $(\sigma_1^k, \sigma_2^k)_{k \in \mathbf{N}}$ converging to some (σ_1, σ_2), and every sequence $(\hat{\sigma}_1^k, \hat{\sigma}_2^k)_{k \in \mathbf{N}}$ converging to some $(\hat{\sigma}_1, \hat{\sigma}_2)$ with $(\hat{\sigma}_1^k, \hat{\sigma}_2^k) \in C^{opt}(\sigma_1^k, \sigma_2^k)$ for every k, it holds that $(\hat{\sigma}_1, \hat{\sigma}_2) \in C^{opt}(\sigma_1, \sigma_2)$.

Suppose, contrary to what we want to prove, that $(\hat{\sigma}_1, \hat{\sigma}_2) \notin C^{opt}(\sigma_1, \sigma_2)$. Then, there is some player i such that $\hat{\sigma}_i$ assigns positive probability to some c_i, whereas c_i is not optimal under $e_i[\sigma_1, \sigma_2]$. But then, if k is large enough, $\hat{\sigma}_i^k$ assigns positive probability to c_i, and c_i is not optimal under $e_i[\sigma_1^k, \sigma_2^k]$. However, this contradicts the assumption that $(\hat{\sigma}_1^k, \hat{\sigma}_2^k) \in C^{opt}(\sigma_1^k, \sigma_2^k)$. So, we conclude that $(\hat{\sigma}_1, \hat{\sigma}_2) \in C^{opt}(\sigma_1, \sigma_2)$, and hence the correspondence C^{opt} is upper-semicontinuous.

Summarizing, we see that the set $A = \Delta(C_1) \times \Delta(C_2)$ is nonempty, compact and convex, and that the correspondence C^{opt} from A to A is upper-semicontinuous and convex-valued. By Kakutani's fixed point theorem (Theorem 4.5.1) it then follows that C^{opt} has at least one fixed point $(\sigma_1^*, \sigma_2^*) \in A$. That is, there is some $(\sigma_1^*, \sigma_2^*) \in A$ with

$$(\sigma_1^*, \sigma_2^*) \in C^{opt}(\sigma_1^*, \sigma_2^*).$$

By definition of C^{opt} this means that for every player i, we have that $\sigma_i^* \in \Delta(C_i^{opt}(\sigma_1^*, \sigma_2^*))$. So, for every player i, the probability distribution σ_i^* only assigns positive probability to choices c_i that are optimal under the induced second-order expectation $e_i[\sigma_1^*, \sigma_2^*]$. This means, however, that (σ_1^*, σ_2^*) is a psychological Nash equilibrium. So, a psychological Nash equilibrium always exists. ■

Proof of Theorem 9.1.4. See the arguments in Section 9.1.4. ■

9.4.2 Proofs of Section 9.2

Proof of Theorem 9.2.1. (a) Suppose first that the belief hierarchy β_i is symmetric and expresses common belief in rationality. Then, we know by Theorem 4.2.1 that the belief hierarchy β_i is induced by a common prior π^* on choice-type combinations. Suppose that, within a beliefs diagram in choice-type representation, β_i starts at the choice-type pair (c_i^*, t_i^*). We say that a choice-type pair (c_j, t_j) can be reached within one step from (c_i^*, t_i^*) if $\pi((c_j, t_j) \mid (c_i^*, t_i^*)) > 0$. Say that a choice-type pair (c_i, t_i) can be reached within two steps from (c_i^*, t_i^*) if there is a pair (c_j, t_j) that can be reached within one step from (c_i^*, t_i^*) such that $\pi((c_i, t_i) \mid (c_j, t_j)) > 0$. For $k \geq 3$, we inductively define reachability within k steps as follows: If k is even, then say that a choice-type pair (c_i, t_i) can be reached within k steps from (c_i^*, t_i^*) if there is a pair (c_j, t_j) that can be reached within $k-1$ steps from (c_i^*, t_i^*) such that $\pi((c_i, t_i) \mid (c_j, t_j)) > 0$. If k is odd, then say that a choice-type pair (c_j, t_j) can be reached within k steps from (c_i^*, t_i^*) if there is a pair (c_i, t_i) that can be reached within $k-1$ steps from (c_i^*, t_i^*) such that $\pi((c_j, t_j) \mid (c_i, t_i)) > 0$.

Let $(C_i \times T_i)^*$ and $(C_j \times T_j)^*$ be the sets of choice-type pairs that can be reached within finitely many steps from (c_i^*, t_i^*). Then, $\pi^*(c_i, t_i) > 0$ for every $(c_i, t_i) \in (C_i \times T_i)^*$ and $\pi^*(c_j, t_j) > 0$ for every $(c_j, t_j) \in (C_j \times T_j)^*$. Moreover, let π be the restriction of π^* to $(C_i \times T_i)^*$ and $(C_j \times T_j)^*$ given by

$$\pi((c_i, t_i), (c_j, t_j)) := \frac{\pi^*((c_i, t_i), (c_j, t_j))}{\sum_{(c_i', t_i') \in (C_i \times T_i)^*,\ (c_j', t_j') \in (C_j \times T_j)^*} \pi^*((c_i', t_i'), (c_j', t_j'))}$$

for every $(c_i, t_i) \in (C_i \times T_i)^*$ and $(c_j, t_j) \in (C_j \times T_j)^*$, and let $\pi((c_i, t_i), (c_j, t_j)) := 0$ otherwise.

Then, it may be verified that the belief hierarchy β_i is induced by the common prior π. We show that π is a psychological correlated equilibrium.

Let T_i^* be the set types that enter in $(C_i \times T_i)^*$, and similarly for player j. Assume, without loss of generality, that for every two choice-type pairs $(c_i, t_i), (c_i', t_i') \in (C_i \times T_i)^*$ with $c_i \neq c_i'$ we have that $t_i \neq t_i'$. Then, for every type $t_i \in T_i^*$ there is a unique choice $c_i[t_i] \in C_i^*$ such that $(c_i[t_i], t_i) \in (C_i \times T_i)^*$. Similarly for player j.

We create an epistemic model with sets of types T_i^* and T_j^*, and where the beliefs of the types are given by

$$b_i(t_i)(c_j, t_j) := \pi((c_j, t_j) \mid (c_i[t_i], t_i)) \tag{9.4.1}$$

for every $t_i \in T_i^*$, and every $(c_j, t_j) \in (C_j \times T_j)^*$. Note that $\pi((c_j, t_j) \mid (c_i[t_i], t_i))$ is well-defined since $\pi^*(c_i[t_i], t_i) > 0$. Similarly for player j's types.

Recall that the belief hierarchy β_i starts at the choice-type pair $(c_i^*, t_i^*) = (c_i[t_i^*], t_i^*)$. Then, by construction, the belief hierarchy β_i is the belief hierarchy induced by the type t_i^* within this epistemic model. We can always select the choice c_i^* such that c_i^* is optimal for t_i^*, as this does not affect the belief hierarchy β_i. Let us therefore assume, without loss of generality, that c_i^* is optimal for t_i^*. In other words, $c_i[t_i^*]$ is optimal for t_i^*.

We will now show that for every $t_i \in T_i{}^*$, the choice $c_i[t_i]$ is optimal for t_i. If $t_i = t_i^*$ then we know this from our assumption above. Assume now that $t_i \neq t_i^*$. Then, $(c_i[t_i], t_i) \in (C_i \times T_i)^*$. In view of (9.4.1), there is a choice-type pair (c_j, t_j) reachable from (c_i^*, t_i^*) such that $b_j(t_j)(c_i[t_i], t_i) > 0$. As the belief hierarchy β_i expresses common belief in rationality, and β_i is the belief hierarchy held by the type t_i^*, we conclude that t_i^* expresses common belief in rationality. Since (c_j, t_j) is reachable from (c_i^*, t_i^*), it follows that t_j believes in i's rationality. As $b_j(t_j)(c_i[t_i], t_i) > 0$, it must thus be that $c_i[t_i]$ is optimal for t_i. In a similar way, it can be shown that $c_j[t_j]$ is optimal for t_j for every type $t_j \in T_j^*$.

Now, take some $(c_i, t_i) \in C_i \times T_i^*$ with $\pi(c_i, t_i) > 0$. Then, $c_i = c_i[t_i]$. By our insights above, we thus know that $c_i[t_i]$ is optimal for t_i. By (9.4.1), the second-order expectation of type t_i is $e_i(\cdot \mid$

$\pi, (c_i, t_i))$. As c_i is optimal for t_i, it follows that c_i is optimal for the induced second-order expectation $e_i(\cdot \mid \pi, (c_i, t_i))$. As the same can be shown for choice-type pairs $(c_j, t_j) \in C_j \times T_j^*$ with $\pi(c_j, t_j) > 0$, we conclude that π is a psychological correlated equilibrium. Hence, the belief hierarchy β_i is induced by a psychological correlated equilibrium.

(b) Assume next that the belief hierarchy β_i is induced by a psychological correlated equilibrium π. As π is a common prior on choice-type combinations, it follows by Theorem 4.2.1 that β_i is symmetric. It remains to show that β_i expresses common belief in rationality.

Suppose that β_i is generated within a beliefs diagram in choice-type representation, and that β_i starts at the choice-type pair (c_i^*, t_i^*). Let $(C_i \times T_i)^*$ and $(C_j \times T_j)^*$ be the sets of choice-type pairs that enter in this beliefs diagram. Moreover, let T_i^* and T_j^* be the sets of types that enter in the beliefs diagram. Similarly to part (a), we assume that for every $t_i \in T_i^*$ there is a unique choice $c_i[t_i]$ such that $(c_i[t_i], t_i) \in (C_i \times T_i)^*$, and similarly for player j.

We construct an epistemic model with sets of types T_i^* and T_j^*, and where the beliefs of the types are given by

$$b_i(t_i)(c_j, t_j) := \pi((c_j, t_j) \mid (c_i[t_i], t_i)) \tag{9.4.2}$$

for every $t_i \in T_i^*$, and every $(c_j, t_j) \in (C_j \times T_j)^*$. Note that $\pi(c_i[t_i], t_i) > 0$ for all $(c_i[t_i], t_i) \in (C_i \times T_i)^*$, and hence $\pi((c_j, t_j) \mid (c_i[t_i], t_i))$ is well-defined. Similarly for player j's types.

Recall that the belief hierarchy β_i is induced by the psychological correlated equilibrium π and starts at the choice-type pair (c_i^*, t_i^*). In view of (9.4.2), the belief hierarchy β_i is precisely the belief hierarchy held by the type t_i^*. We will now show that t_i^* expresses common belief in rationality. For this, it is sufficient to show that every type in the epistemic model above believes in the opponent's rationality.

Take a type $t_i \in T_i^*$ and a choice-type pair $(c_j, t_j) \in C_j \times T_j^*$ with $b_i(t_i)(c_j, t_j) > 0$. Then, we know by (9.4.2) that $\pi((c_j, t_j) \mid (c_i[t_i], t_i)) > 0$. This implies that $\pi(c_j, t_j) > 0$. As π is a psychological correlated equilibrium, we know that c_j is optimal for the induced second-order expectation $e_j(\cdot \mid \pi, (c_j, t_j))$. By (9.4.2) we know that t_j's second-order expectation is $e_j(\cdot \mid \pi, (c_j, t_j))$. Therefore, c_j is optimal for the type t_j. We thus conclude that t_i believes in j's rationality. In a similar fashion, it can be shown that every type $t_j \in T_j^*$ believes in i's rationality.

As such, every type in the epistemic model believes in the opponent's rationality. This, in turn, implies that every type expresses common belief in rationality. In particular, type t_i^* expresses common belief in rationality, which means that belief hierarchy β_i expresses common belief in rationality. This completes the proof. ■

Proof of Theorem 9.2.2. Follows from the arguments in Section 9.2.2. ■

Proof of Theorem 9.2.3. Follows from the arguments in Section 9.2.2. ■

Proof of Theorem 9.2.4. Follows from the arguments in Section 9.2.3. ■

Proof of Theorem 9.2.5. Follows from the arguments in Section 9.2.3. ■

Proof of Theorem 9.2.3. Follows from the arguments in Section 9.2.3. ■

Solutions to In-Chapter Questions

Question 9.1.1. The induced second-order expectation $e_1[\sigma_1, \sigma_2]$ for you is given by

$$\begin{aligned} e_1[\sigma_1, \sigma_2] &= (0.6) \cdot (0.2) \cdot (n, n) + (0.4) \cdot (0.2) \cdot (b, n) + (0.6) \cdot (0.5) \cdot (n, r) \\ &\quad + (0.4) \cdot (0.5) \cdot (b, r) + (0.6) \cdot (0.3) \cdot (n, b) + (0.4) \cdot (0.3) \cdot (b, b) \\ &= (0.12) \cdot (n, n) + (0.08) \cdot (b, n) + (0.3) \cdot (n, r) \\ &\quad + (0.2) \cdot (b, r) + (0.18) \cdot (n, b) + (0.12) \cdot (b, b), \end{aligned}$$

whereas the induced second-order expectation $e_2[\sigma_1, \sigma_2]$ for Barbara is given by

$$\begin{aligned} e_2[\sigma_1, \sigma_2] &= (0.2) \cdot (0.6) \cdot (n, n) + (0.5) \cdot (0.6) \cdot (r, n) + (0.3) \cdot (0.6) \cdot (b, n) \\ &\quad + (0.2) \cdot (0.4) \cdot (n, b) + (0.5) \cdot (0.4) \cdot (r, b) + (0.3) \cdot (0.4) \cdot (b, b) \\ &= (0.12) \cdot (n, n) + (0.3) \cdot (r, n) + (0.18) \cdot (b, n) \\ &\quad + (0.08) \cdot (n, b) + (0.2) \cdot (r, b) + (0.12) \cdot (b, b). \end{aligned}$$

Question 9.1.2. Suppose that i believes that j assigns a positive probability to i's choice c_i. As i's second-order belief is given by σ_i, we conclude that $\sigma_i(c_i) > 0$. By the definition of a psychological Nash equilibrium, choice c_i is optimal for the induced second-order expectation $e_i[\sigma_1, \sigma_2]$. Since the simple belief hierarchy is generated by (σ_1, σ_2), player i believes that j believes that i's second-order expectation is $e_i[\sigma_1, \sigma_2]$. Hence, i believes that j assigns a positive probability to i's choice c_i only if c_i is optimal for player i, given what i believes that j believes about i's second-order expectation. In other words, i believes that j believes in i's rationality.

Question 9.1.3. Suppose, on the contrary, that $\sigma_1(\textit{bracelet}) > 0$. Then, *bracelet* must be optimal for you under the second-order expectation $e_1[\sigma_1, \sigma_2]$. However, we have seen in Section 8.6 that your choice *bracelet* is strictly dominated in your decision problem, and therefore not optimal for any second-order expectation. Thus, $\sigma_1(\textit{bracelet}) = 0$.

Assume, on the contrary, that $\sigma_2(\textit{bracelet}) > 0$. Then, *bracelet* must be optimal for Barbara under the second-order expectation $e_2[\sigma_1, \sigma_2]$. Since Barbara's conditional preference relation only depends on her first-order belief, we must have that *bracelet* is optimal for Barbara under the first-order belief σ_1. However, since $\sigma_1(\textit{bracelet}) = 0$, Barbara's choice *bracelet* can never be optimal under the first-order belief σ_1. We thus conclude that $\sigma_2(\textit{bracelet}) = 0$.

Question 9.2.1. By looking at the beliefs diagram in Figure 9.2.1 we see that the second-order expectation $e_2(\cdot \mid (\textit{white}_2,\ t_2^w))$ is given by

$$\begin{aligned} e_2((\textit{black}_1, t_1^b),\ (\textit{black}_2,\ t_2^b) \mid (\textit{white}_2,\ t_2^w)) &= \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}, \\ e_2((\textit{black}_1, t_1^b),\ (\textit{white}_2,\ t_2^w) \mid (\textit{white}_2,\ t_2^w)) &= \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}, \\ e_2((\textit{white}_1, t_1^w),\ (\textit{black}_2,\ t_2^b) \mid (\textit{white}_2,\ t_2^w)) &= \tfrac{1}{2} \cdot \tfrac{2}{3} = \tfrac{1}{3} \text{ and} \\ e_2((\textit{white}_1, t_1^w),\ (\textit{white}_2,\ t_2^w) \mid (\textit{white}_2,\ t_2^w)) &= \tfrac{1}{2} \cdot \tfrac{1}{3} = \tfrac{1}{6}. \end{aligned}$$

The expected utilities for Barbara of choosing $\textit{black}_2$ and $\textit{white}_2$ are thus given by

$$\begin{aligned} u_2(\textit{black}_2,\ e_2(\cdot \mid (\textit{white}_2,\ t_2^w)) &= 2 \text{ and} \\ u_2(\textit{white}_2,\ e_2(\cdot \mid (\textit{white}_2,\ t_2^w)) &= \tfrac{1}{4} \cdot 8 + \tfrac{1}{4} \cdot 0 + \tfrac{1}{3} \cdot 0 + \tfrac{1}{6} \cdot 0 = 2, \end{aligned}$$

which implies that $white_2$ (but also $black_2$) is optimal for $e_2(\cdot \mid (white_2,\, t_2^w))$.

Question 9.2.2. The induced second-order expectations are

$$\begin{aligned} e_1[\sigma_1, \sigma_2] &= (0.25) \cdot (black_2,\ black_1) + (0.25) \cdot (black_2,\ white_1) \\ &\quad +(0.25) \cdot (white_2,\ black_1) + (0.25) \cdot (white_2,\ white_1) \end{aligned}$$

and

$$\begin{aligned} e_2[\sigma_1, \sigma_2] &= (0.25) \cdot (black_1,\ black_2) + (0.25) \cdot (black_1,\ white_2) \\ &\quad +(0.25) \cdot (white_1,\ black_2) + (0.25) \cdot (white_1,\ white_2). \end{aligned}$$

For you, the expected utilities of choosing *black* or *white* under the second-order expectation $e_1[\sigma_1, \sigma_2]$ are both 2. Hence, both *black* and *white* are optimal for you under the second-order expectation $e_1[\sigma_1, \sigma_2]$. Similarly, both *black* and *white* are optimal for Barbara under the second-order expectation $e_2[\sigma_1, \sigma_2]$. This implies that (σ_1, σ_2) is a psychological Nash equilibrium.

Question 9.2.3. To make the calculations more compact, we denote the colors by b, g and r. Consider your choice b_1 with $\hat{\pi}(b_1) > 0$. In the induced second-order expectation $e_1(\cdot \mid \hat{\pi}, b_1)$ we have that

$$\begin{aligned} e_1((b_2, b_1) \mid \hat{\pi}, b_1) &= \hat{\pi}(b_2 \mid b_1) \cdot \hat{\pi}(b_1 \mid b_2) = \frac{\hat{\pi}(b_1, b_2)}{\hat{\pi}(b_1)} \cdot \frac{\hat{\pi}(b_1, b_2)}{\hat{\pi}(b_2)} \\ &= \frac{1/130}{101/130} \cdot \frac{1/130}{5/130} = \frac{1}{101} \cdot \frac{1}{5} = \frac{1}{505}. \end{aligned}$$

In a similar way it can be verified that

$$\begin{aligned} e_1((b_2, g_1) \mid \hat{\pi}, b_1) &= \frac{1}{101} \cdot \frac{4}{5} = \frac{4}{505}, \\ e_1((r_2, b_1) \mid \hat{\pi}, b_1) &= \frac{100}{101} \cdot \frac{4}{5} = \frac{400}{505} \text{ and} \\ e_1((r_2, g_1) \mid \hat{\pi}, b_1) &= \frac{100}{101} \cdot \frac{1}{5} = \frac{100}{505}. \end{aligned}$$

Thus, the expected utilities of choosing b_1 and g_1 under $e_1(\cdot \mid \hat{\pi}, b_1)$ are

$$u_1(b_1,\ e_1(\cdot \mid \hat{\pi}, b_1)) = \frac{1}{505} \cdot 0 + \frac{4}{505} \cdot 3 + \frac{400}{505} \cdot 3 + \frac{100}{505} \cdot 6 = \frac{1812}{505}$$

and

$$u_1(g_1,\ e_1(\cdot \mid \hat{\pi}, b_1)) = \frac{1}{505} \cdot 4 + \frac{4}{505} \cdot 2 + \frac{400}{505} \cdot 4 + \frac{100}{505} \cdot 2 = \frac{1812}{505}.$$

Hence, $u_1(b_1, e_1(\cdot \mid \hat{\pi}, b_1)) = u_1(g_1, e_1(\cdot \mid \hat{\pi}, b_1))$, which means that, in particular, $u_1(b_1, e_1(\cdot \mid \hat{\pi}, b_1)) \geq u_1(g_1, e_1(\cdot \mid \hat{\pi}, b_1))$.

Next, consider your choice g_1 with $\hat{\pi}(g_1) > 0$. The induced second-order expectation $e_1(\cdot \mid \hat{\pi}, g_1)$ is given by

$$\begin{aligned} e_1((b_2, b_1) \mid \hat{\pi}, g_1) &= \frac{4}{29} \cdot \frac{1}{5} = \frac{4}{145} \\ e_1((b_2, g_1) \mid \hat{\pi}, g_1) &= \frac{4}{29} \cdot \frac{4}{5} = \frac{16}{145}, \\ e_1((r_2, b_1) \mid \hat{\pi}, g_1) &= \frac{25}{29} \cdot \frac{4}{5} = \frac{100}{145} \text{ and} \\ e_1((r_2, g_1) \mid \hat{\pi}, g_1) &= \frac{25}{29} \cdot \frac{1}{5} = \frac{25}{145}. \end{aligned}$$

Thus, the expected utilities of choosing g_1 and b_1 under $e_1(\cdot \mid \hat{\pi}, g_1)$ are

$$u_1(g_1,\ e_1(\cdot \mid \hat{\pi}, g_1)) = \frac{4}{145} \cdot 4 + \frac{16}{145} \cdot 2 + \frac{100}{145} \cdot 4 + \frac{25}{145} \cdot 2 = \frac{498}{145}$$

and

$$u_1(b_1,\ e_1(\cdot \mid \hat{\pi}, g_1)) = \frac{4}{145} \cdot 0 + \frac{16}{145} \cdot 3 + \frac{100}{145} \cdot 3 + \frac{25}{145} \cdot 6 = \frac{498}{145}.$$

Hence, $u_1(g_1,\ e_1(\cdot \mid \hat{\pi}, g_1)) = u_1(b_1,\ e_1(\cdot \mid \hat{\pi}, g_1))$, which means that, in particular, $u_1(g_1,\ e_1(\cdot \mid \hat{\pi}, g_1)) \geq u_1(b_1,\ e_1(\cdot \mid \hat{\pi}, g_1))$.

Now, consider Barbara's choice b_2 with $\hat{\pi}(b_2) > 0$. The induced second-order expectation $e_2(\cdot \mid \hat{\pi}, b_2)$ is given by

$$\begin{aligned}
e_2((b_1, b_2) \mid \hat{\pi}, b_2) &= \frac{1}{5} \cdot \frac{1}{101} = \frac{1}{505}, \\
e_2((b_1, r_2) \mid \hat{\pi}, b_2) &= \frac{1}{5} \cdot \frac{100}{101} = \frac{100}{505}, \\
e_2((g_1, b_2) \mid \hat{\pi}, b_2) &= \frac{4}{5} \cdot \frac{4}{29} = \frac{16}{145} \text{ and} \\
e_2((g_1, r_2) \mid \hat{\pi}, b_2) &= \frac{4}{5} \cdot \frac{25}{29} = \frac{100}{145}.
\end{aligned}$$

Thus, the expected utilities of choosing b_2 and r_2 under $e_2(\cdot \mid \hat{\pi}, b_2)$ are

$$u_2(b_2,\ e_2(\cdot \mid \hat{\pi}, b_2)) = \frac{1}{505} \cdot 0 + \frac{100}{505} \cdot 2 + \frac{16}{145} \cdot 2 + \frac{100}{145} \cdot 4 \approx 3.38.$$

and

$$u_2(r_2,\ e_2(\cdot \mid \hat{\pi}, b_2)) = \frac{1}{505} \cdot 6 + \frac{100}{505} \cdot 3 + \frac{16}{145} \cdot 6 + \frac{100}{145} \cdot 3 \approx 3.34.$$

Hence, $u_2(b_2,\ e_2(\cdot \mid \hat{\pi}, b_2)) \geq u_2(r_2,\ e_2(\cdot \mid \hat{\pi}, b_2))$.

We finally turn to Barbara's choice r_2 with $\hat{\pi}(r_2) > 0$. The induced second-order expectation $e_2(\cdot \mid \hat{\pi}, r_2)$ is given by

$$\begin{aligned}
e_2((b_1, b_2) \mid \hat{\pi}, r_2) &= \frac{4}{5} \cdot \frac{1}{101} = \frac{4}{505}, \\
e_2((b_1, r_2) \mid \hat{\pi}, r_2) &= \frac{4}{5} \cdot \frac{100}{101} = \frac{400}{505}, \\
e_2((g_1, b_2) \mid \hat{\pi}, r_2) &= \frac{1}{5} \cdot \frac{4}{29} = \frac{4}{145} \text{ and} \\
e_2((g_1, r_2) \mid \hat{\pi}, r_2) &= \frac{1}{5} \cdot \frac{25}{29} = \frac{25}{145}.
\end{aligned}$$

Thus, the expected utilities of choosing r_2 and b_2 under $e_2(\cdot \mid \hat{\pi}, r_2)$ are

$$u_2(r_2,\ e_2(\cdot \mid \hat{\pi}, r_2)) = \frac{4}{505} \cdot 6 + \frac{400}{505} \cdot 3 + \frac{4}{145} \cdot 6 + \frac{25}{145} \cdot 3 \approx 3.11.$$

and

$$u_2(b_2,\ e_2(\cdot \mid \hat{\pi}, r_2)) = \frac{4}{505} \cdot 0 + \frac{400}{505} \cdot 2 + \frac{4}{145} \cdot 2 + \frac{25}{145} \cdot 4 \approx 2.33.$$

Hence, $u_2(r_2,\ e_2(\cdot \mid \hat{\pi}, r_2)) \geq u_2(b_2,\ e_2(\cdot \mid \hat{\pi}, r_2))$.

In view of all this, we conclude that $\hat{\pi}$ is a canonical psychological correlated equilibrium.

Problems

Problem 9.1: High jump.

In a few weeks from now there will be the regional athletics championships. Both Barbara and you would like to participate, and high jump is the favorite discipline for the two of you. From experience, you know that by practicing very hard you will be able to jump 2.20 meters, whereas you can easily jump 1.80 meters with very little practice. The question is: What height will you practice for during the next few weeks?

Suppose you can choose to practice for 1.80 meters, 1.90 meters, 2.00 meters, 2.10 meters or 2.20 meters. Clearly, jumping higher requires more practice. More precisely, if you aim to jump a height of h, then your mental and physical cost of practicing for that height will be $h^2/400$, where the height h is expressed in centimeters.

Moreover, you would like to meet, or even exceed, Barbara's expectations: If you jump a height of h, and Barbara believes that you will jump a height of h', then your mental bonus of meeting, or exceeding, Barbara's expectation will be 2 if $h \geq h'$ and 0 otherwise.

Your utility is given by the height h you aim to jump (expressed in centimeters), minus the cost of practicing for the height of h, plus, possibly, the mental bonus from meeting, or exceeding, Barbara's expectation. The choices and the utilities for Barbara are similar to yours.

(a) Model this situation as a psychological game by writing down the decision problem for you. The decision problem for Barbara is similar, by symmetry.

(b) Perform the iterated elimination of choices and states. Which heights survive for you and Barbara? Can you guarantee, on the basis of Theorem 8.6.1, that this procedure yields precisely those choices you can rationally make under common belief in rationality? Explain your answer.

(c) Create a beliefs diagram, with solid arrows only, that uses all the heights for you and Barbara that have survived the procedure in (b). Translate this beliefs diagram into an epistemic model, and explain why all types express common belief in rationality.

(d) Find the heights that you can rationally practice for under common belief in rationality with a simple belief hierarchy.

(e) Consider the beliefs diagram in Figure 9.4.1. Show that all belief hierarchies in the beliefs diagram are symmetric, by finding a symmetric *weighted* beliefs diagram that induces this beliefs diagram.

(f) Translate the symmetric weighted beliefs diagram from (e) into a common prior on choice combinations, and show that this common prior is a canonical psychological correlated equilibrium.

Throughout the week, your preferences have changed: From now on, you only care about *exceeding* Barbara's preferences. That is, if you jump a height of h, and Barbara believes that you would jump a hight of h', then the mental bonus from exceeding Barbara's preferences is 2 if $h > h'$, and 0 otherwise. Apart from this, your utilities are built up in the same way as above. Similarly for Barbara.

(g) Model this new situation as a psychological game, by writing down the decision problem for you.

(h) Which heights can you rationally aim for under common belief in rationality?

(i) Create a beliefs diagram, with solid arrows only, that uses all the heights for you and Barbara that you found in (h). Under common belief in rationality, what is the highest probability by which you can believe to exceed Barbara's expectations?

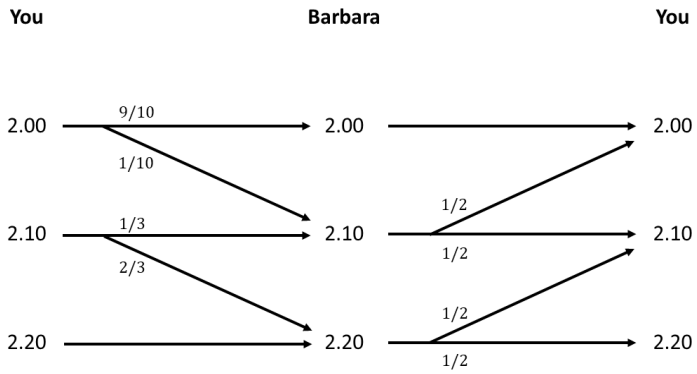

Figure 9.4.1 Beliefs diagram for Problem 9.1 (e)

*(j) Find the unique psychological Nash equilibrium. What heights can you rationally aim for under common belief in rationality with a simple belief hierarchy? Under common belief in rationality with a simple belief hierarchy, what is the highest probability by which you can believe to exceed Barbara's expectations?

Problem 9.2: The Cooper test.

As a preparation for the athletics championships, you and Barbara have hired Chris as a personal coach. The first thing Chris proposes is that you both do the Cooper test: Within 12 minutes you must run the largest possible distance. You and Barbara know that, in principle, you would both be able to run a distance of 2300 meters, 2500 meters or 2700 meters. The question is: Which of these three distances are you and Barbara planning to run?

Obviously, running a larger distance comes at a larger physical and mental cost. Suppose that the cost of running d meters is $d^2/5000$.

You care both about beating Barbara and exceeding Barbara's expectations. More precisely, if you run a distance of d_1, Barbara runs a distance of d_2, and Barbara believes that you run a distance of d_1', then your mental bonus of beating Barbara is $0.1 \cdot (d_1 - d_2)$ if $d_1 > d_2$ and 0 otherwise, and your mental bonus of exceeding Barbara's expectation is $0.1 \cdot (d_1 - d_1')$ if $d_1 > d_1'$ and 0 otherwise.

On the other hand, Barbara's only cares about exceeding your expectation. Her mental bonus for exceeding your expectation is similarly to yours.

Your utility is given by the distance you run in meters, minus the cost of running this distance, plus (possibly) the mental bonus from beating Barbara, plus (possibly) the mental bonus from exceeding Barbara's expectation. Barbara's utility is the distance she runs in meters, minus the cost of running this distance, plus (possibly) the mental bonus from exceeding your expectation.

(a) Do the preferences over your choices depend only on your first-order belief, only on your second-order belief, or on both? Answer the same question for Barbara.

(b) Set up the decision problems for you and Barbara.

(c) Perform the iterated elimination of choices and states. Which distances survive for you and Barbara? Can you guarantee, on the basis of Theorem 8.6.1, that this procedure yields precisely those choices you can rationally make under common belief in rationality? Explain your answer.

(d) Show that each of the distances for you and Barbara found in (c) can indeed rationally be chosen under common belief in rationality. To this purpose, construct a beliefs diagram with solid arrows only that contains all the choices found in (c). What is the highest probability by which you believe you can beat Barbara under common belief in rationality? What is the highest probability by which you believe you can exceed Barbara's expectations under common belief in rationality? Translate this beliefs diagram into an epistemic model, and explain why all types express common belief in rationality.

***(e)** Find the unique psychological Nash equilibrium. What distances can you and Barbara rationally choose under common belief in rationality with a simple belief hierarchy? What is the highest probability by which you believe you can beat Barbara under common belief in rationality with a simple belief hierarchy? What is the highest probability by which you believe you can exceed Barbara's expectations under common belief in rationality with a simple belief hierarchy?

(f) What distances can you and Barbara rationally choose under common belief in rationality with a symmetric belief hierarchy? What distances can you and Barbara rationally choose under common belief in rationality with a symmetric belief hierarchy that uses one theory per choice? Explain your answer.

Problem 9.3: How many disciplines?

Remember from Problem 9.1 that soon there will be the yearly regional athletics championships. Both you and Barbara want to participate, and now it is time to decide at how many different disciplines you would like to compete. According to the rules, you can subscribe for at most four disciplines. Needless to say that you must subscribe for at least one discipline. Of course, the more disciplines you choose, the more you will have to practice. Suppose that the mental and physical cost of practicing for d disciplines is d^2, and similarly for Barbara.

The utility you derive from participating at the championship is 4. Moreover, you would like to exceed Barbara's expectations by the number of disciplines you choose. More precisely, if you choose d_1 disciplines, and Barbara believes that you choose d_1' disciplines, then the mental bonus of exceeding Barbara's expectations is 4 if $d_1 > d_1'$, and it is 0 otherwise.

Your total utility is 4 from participating at the championship, minus the cost of practicing, plus (possibly) the bonus from exceeding Barbara's expectation.

Also for Barbara, the utility she derives from participating at the championship is 4. However, rather than trying to exceed your expectations, Barbara would like to choose more disciplines than you do. More concretely, if she chooses more disciplines than you do, then her mental bonus will be 6. Barbara's total utility is therefore 4 from participating at the championship, minus the cost of practicing, plus (possibly) the bonus from choosing more disciplines than you do.

(a) Model this story as a psychological game, by specifying the decision problems for you and Barbara.

(b) Suppose we would like to find the choices you can rationally make under common belief in rationality. Explain, on the basis of Theorem 8.6.1, why it is sufficient to apply the iterated elimination of choices and states.

(c) How many disciplines can you and Barbara rationally choose under common belief in rationality?

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

Explain why under common belief in rationality, you can expect to exceed Barbara's expectations with probability 1.

(e) Find the unique psychological Nash equilibrium. How many disciplines can you and Barbara rationally choose under common belief in rationality with a simple belief hierarchy?

(f) Suppose your belief hierarchy is simple and expresses common belief in rationality. What, then, is the unique probability by which you can expect to exceed Barbara's expectations?

(g) How many different disciplines can you rationally choose under common belief in rationality with a symmetric belief hierarchy using one theory per choice?

***(h)** Show that under common belief in rationality with a symmetric belief hierarchy that uses one theory per choice, you can only expect to exceed Barbara's expectations with probability 3/4.

(**Hint to (h):** Consider a canonical psychological correlated equilibrium $\hat{\pi}$. Show the following steps.

Step 1. Show that $\hat{\pi}(1_1, 2_2) > 0$.

Step 2. Show that $\hat{\pi}(2_1, 2_2) > 0$.

Step 3. Show that $\hat{\pi}(1_1, 2_2) > 0$, $\hat{\pi}(2_1, 2_2) > 0$ and $\hat{\pi}(2_1, 1_2) = 0$ is only possible if $\hat{\pi}(1_1, 2_2) = 3/4$, $\hat{\pi}(2_1, 2_2) = 1/4$, $\hat{\pi}(1_1, 1_2) = 0$ and $\hat{\pi}(2_1, 1_2) = 0$. In this case, conclude that you expect to exceed Barbara's expectation with probability 3/4 by choosing 2 disciplines.

Step 4. Show that $\hat{\pi}(1_1, 2_2) > 0$, $\hat{\pi}(2_1, 2_2) > 0$ and $\hat{\pi}(2_1, 1_2) > 0$ is impossible. To do so, assume that $\hat{\pi}(1_1, 2_2) > 0$, $\hat{\pi}(2_1, 2_2) > 0$ and $\hat{\pi}(2_1, 1_2) > 0$ and prove the following substeps:

Step 4.1. Show that $3 \geq 4 \cdot [\hat{\pi}(1_2 \mid 1_1) \cdot \hat{\pi}(1_1 \mid 1_2) + \hat{\pi}(2_2 \mid 1_1) \cdot \hat{\pi}(1_1 \mid 2_2)]$.

Step 4.2. Based on Step 4.1, show that $4 \cdot \hat{\pi}(1_2 \mid 1_1) \cdot [\hat{\pi}(1_1 \mid 2_2) - \hat{\pi}(1_1 \mid 1_2)] \geq 4 \cdot \hat{\pi}(1_1 \mid 2_2) - 3$.

Step 4.3. Show that $3 \leq 4 \cdot [\hat{\pi}(1_2 \mid 2_1) \cdot \hat{\pi}(1_1 \mid 1_2) + \hat{\pi}(2_2 \mid 2_1) \cdot \hat{\pi}(1_1 \mid 2_2)]$.

Step 4.4. Based on Step 4.3, show that $4 \cdot \hat{\pi}(1_2 \mid 2_1) \cdot [\hat{\pi}(1_1 \mid 2_2) - \hat{\pi}(1_1 \mid 1_2)] \leq 4 \cdot \hat{\pi}(1_1 \mid 2_2) - 3$.

Step 4.5. Show that $\hat{\pi}(1_1 \mid 1_2) \leq 1/2$.

Step 4.6. Based on Steps 4.3 and 4.5, show that $\hat{\pi}(1_1 \mid 2_2) \geq 3/4$.

Step 4.7. On the basis of Steps 4.2, 4.4, 4.5 and 4.6, show that $\hat{\pi}(1_2 \mid 1_1) \geq \hat{\pi}(1_2 \mid 2_1)$.

Step 4.8. Of the basis of Step 4.7, show that $\hat{\pi}(1_1 \mid 1_2) \geq \hat{\pi}(1_1 \mid 2_2)$. This, however, is a contradiction to Steps 4.5 and 4.6. Hence, it follows by Step 3 that you expect to exceed Barbara's expectation with probability 3/4 by choosing 2 disciplines.)

Literature

Psychological Nash equilibrium and psychological correlated equilibrium. The concept of psychological Nash equilibrium has been introduced by Geanakoplos, Pearce and Stacchetti (1989). Their definition is equivalent to ours, although they do not use the label of *simple* belief hierarchies. To the best of our knowledge, the concepts of psychological correlated equilibrium and canonical psychological correlated equilibrium as proposed in this chapter are new.

In addition, Geanakoplos, Pearce and Stacchetti (1989) introduce equilibrium concepts for *dynamic* psychological games, such as subgame perfect psychological equilibrium, trembling hand perfect psychological equilibrium and sequential psychological equilibrium. Also Battigalli and Dufwenberg (2009) consider a version of sequential psychological equilibrium which they call sequential equilibrium.

How psychological Nash equilibrium limits surprise. We have seen in several examples in this chapter that insisting on a *simple* belief hierarchy in combination with common belief in rationality – that is, insisting on psychological Nash equilibrium – can severely limit the ability to surprise the other person. See, for instance, the examples "Barbara's birthday" and "Surprising Barbara".

Mourmans (2017) shows that this is also true for the famous *surprise exam paradox,* where a teacher tries to surprise the student by the day on which he poses the exam. Indeed, if we merely use common belief in rationality without insisting on a simple belief hierarchy, then Mourmans (2017) shows that the teacher is able to fully surprise the student. However, if we additionally insist on a simple belief hierarchy, then it is shown that the teacher is only able to *partially* surprise the student, that is, to surprise the student with a probability that is substantially less than 1.

Geanakoplos (1996) comes to a similar conclusion if we model the *hangman paradox,* which is similar to the surprise exam paradox, as a psychological game and subsequently use the concept of psychological Nash equilibrium. In that case, the judge can at best only partially surprise the prisoner by the day on which the sentence will be executed. The reason, as above, is that the simple belief hierarchy, in combination with common belief in rationality, severely limits the ability of surprising the other person.