Chapter 3

Common Belief in Rationality

3.1 Games as Decision Problems

In Chapter 2 we have investigated decision problems under uncertainty, where the consequence of a choice depends on events that are beyond your control. Such events have been summarized by *states.* From now on we will concentrate on scenarios where the states involve *choices of other people.* These scenarios will be called *games.* We will see that, by reasoning about the rationality of other people, you can possibly eliminate some states, since these involve irrational choices by others. As will become clear throughout this book, reasoning about the decision problems and rationality of other people is really at the core of game theory.

3.1.1 When States Involve Choices of Others

In the leading example of Chapter 2, the events about which you were uncertain concerned the state of the weather. Indeed, which location you would choose for your birthday party heavily depended on your belief about the state of the weather. Without any weather forecast, or any other piece of information that could give an indication about the likely state of the weather, every probabilistic belief about the three states *rainy, stormy* and *calm* would, in principle, be reasonable.

This may no longer be the case, however, if the states involve choices by other people. If you reason about the decision problems and rationality of these other people, then possibly some states may be discarded, as they involve *irrational* choices by others. Consider, for instance, the following example.

Example 3.1: Going to a party.

You and your friend Barbara have been invited to a party this evening. The problem is: What color do you wear? Suppose you only have *blue, green, red* and *yellow* outfits in your wardrobe, and that the same is true for Barbara. Of course, there are some colors you like more and other colors you like

You	*blue*	*green*	*red*	*yellow*
blue	0	4	4	4
green	3	0	3	3
red	2	2	0	2
yellow	1	1	1	0

Barbara	*blue*	*green*	*red*	*yellow*
blue	0	2	2	2
green	1	0	1	1
red	4	4	0	4
yellow	3	3	3	0

Table 3.1.1 Decision problems for "Going to a party"

less. It turns out that, in principle, you prefer *blue* to *green,* you prefer *green* to *red,* and you prefer *red* to *yellow.*

However, you really dislike arriving at the party wearing the same color as Barbara. In that case you would be very disappointed. If you believe, for instance, that Barbara will wear your favorite color *blue,* then you would rather wear *yellow* than *blue.* The problem is that you are *uncertain* about Barbara's color choice.

In this scenario, your possible choices are thus *blue, green, red* and *yellow,* whereas the states correspond to Barbara's choices *blue, green, red* and *yellow.* Assume that your conditional preference relation has the expected utility representation shown in the left-hand panel of Table 3.1.1. Hence, the left-hand table summarizes your decision problem $(C, S, \succsim)$, where C is your set of choices {*blue, green, red, yellow*}, the set of states S contains Barbara's choices {*blue, green, red, yellow*}, and your conditional preference relation $\succsim$ is given by the expected utility representation in that table.

Note that at the state *green,* that is, when Barbara wears *green,* the intensity by which you prefer *blue* to *red* is twice the intensity by which you prefer *red* to *yellow.* Indeed, the utility differences between *blue* and *red* and between *red* and *yellow* are 2 and 1, respectively, and we have seen in Chapter 2 that the expected utility difference between two choices serves as a measure for the intensity by which you prefer one choice to the other. Moreover, at state *green* the intensity by which you prefer *red* to *yellow* is the same as the intensity by which you prefer *yellow* to *green.* In particular, *green* is your least preferred choice if you believe Barbara wears *green,* since you strongly dislike wearing the same color as Barbara. Similar properties hold at the other three states.

Recall from Chapter 2 that a choice is called *rational* if it is optimal for at least one belief. Now, which colors are rational for you and which are not? Clearly, *blue* is optimal for the belief that assigns probability 1 to Barbara wearing *green,* whereas *green* is optimal for the belief that assigns probability 1 to Barbara wearing *blue.*

Question 3.1.1 *Consider the belief* $(0.6)\cdot$ *blue* $+ (0.4)\cdot$ *green, that assigns probability* 0.6 *to Barbara wearing blue, and probability* 0.4 *to Barbara wearing green. What is your preference relation over colors at that belief ?*

In particular, we see that your choice *red* is optimal for this belief. But what about your choice *yellow* ? It turns out that this choice is strictly dominated by a randomized choice.

Question 3.1.2 *Find a randomized choice that strictly dominates your choice yellow for the expected utility representation* u *in Table 3.1.1. Does it strictly dominate yellow for every expected utility representation of* $\succsim$ *?*

By Theorem 2.7.1 we thus conclude that your choice *yellow* is never optimal for any belief about Barbara's choice, and is thus irrational. Summarizing, we see that your choices *blue, green* and *red* are rational, whereas your choice *yellow* is irrational.

3.1.2 Games

But does this automatically mean that your choices *blue, green* and *red* are also plausible? Not necessarily. The reason is that Barbara also holds a conditional preference relation over her choices, and it may well be that some color is irrational for Barbara. But if some color is irrational for Barbara, then it seems natural to assign probability 0 to that color, as you expect Barbara not to choose that particular color. Such reasoning would impose restrictions on your beliefs, which may possibly rule out some of the rational choices above.

To see how this works, suppose that Barbara's preferences over colors are different from yours: In principle, she prefers *red* to *yellow, yellow* to *blue,* and *blue* to *green.* Similarly to you, also Barbara dislikes wearing the same color as you. This leads to the utility function in the right-hand panel of Table 3.1.1, which represents Barbara's conditional preference relation over the four colors. In that table, we have put Barbara's choices in the rows, and your choices in the columns. The reason is that from Barbara's perspective, the *states* of her decision problem consist of *your* choices. Indeed, she is uncertain about your choice of color.

These two decision problems together – one for you and one for Barbara – constitute a *game.* The two decision makers, you and Barbara in this case, are called *players.* Sometimes, we number the players by $1, 2, ...$. Important is that the states in player i's decision problem involve the choices of other players, because player i is uncertain about the choices of other players. At the same time, the choices of i's opponents influence the consequences of i's choices, and thus shape the preference relation that player i has about his own choices. In the example above, the states in your decision problem are Barbara's choices, and the states in Barbara's decision problem are your own choices.

For the formal definition of a game below, we use the following pieces of notation. By C_i we denote the set of choices for player i, whereas C_{-i} denotes the set of choice combinations of players other than i. If there are only two players, as in the example above, then there is only one other player, say j, and hence C_{-i} denotes the set of player j's choices. If there are more than two players, then there is more than one opponent for player i, and C_{-i} contains the choice combinations of two or more opponents. This will be illustrated later, when we study an example with three players.

Definition 3.1.1 (Game) *A **game** specifies*

(a) a finite set of players I,

(b) for every player i a finite set of choices C_i,

(c) for every player i a decision problem $(C_i, C_{-i}, \succsim_i)$, where the set of choices is C_i, the set of states is the set C_{-i} of opponents' choice combinations, and $\succsim_i$ is a conditional preference relation for player i, assigning to every belief $p_i \in \Delta(C_{-i})$ about the opponents' choice combinations a preference relation $\succsim_{i,p_i}$ over i's choices.

Throughout the remainder of this book, we will assume that the players' conditional preference relations $\succsim_i$ will always have an expected utility representation u_i. As such, their decision problems can be summarized by (C_i, C_{-i}, u_i), where u_i is a utility function that assigns to every choice $c_i \in C_i$ and state $c_{-i} \in C_{-i}$ some utility $u_i(c_i, c_{-i})$. By doing so, we thus assume that the conditional preference relations of all players satisfy each of the axioms from Theorem 2.6.1.

The two panels in Table 3.1.1 thus constitute a game, where the set of players is $I = \{$*you, Barbara*$\}$ or, as an abbreviation, $I = \{1, 2\}$, where *you* are player 1 and *Barbara* is player 2. The sets of choices are $C_1 = \{$*blue, green, red, yellow*$\}$ and $C_2 = \{$*blue, green, red, yellow*$\}$, and the players' decision problems are summarized by the expected utility representations in the two panels of Table 3.1.1.

You	*blue*	*red*	*yellow*
blue	0	4	4
green	3	3	3
red	2	0	2
yellow	1	1	0

Table 3.1.2 Reduced decision problem for "Going to a party"

3.1.3 Reasoning about Others' Decision Problems

Recall from above that your choices *blue, green* and *red* are all rational, and that your choice *yellow* is not. We now turn to the question: Which of these rational choices are also *plausible,* in the light of Barbara's decision problem in Table 3.1.1? To answer this question, we must start to *reason* about Barbara's decision problem.

Note first the similarity between Barbara's conditional preference relation and yours: Although her preferences over colors are different from yours, many of the characteristics carry over from your conditional preference relation to hers. Similarly to the analysis of your decision problem, it can be shown that for Barbara, her choices *red, yellow* and *blue* are rational, whereas her choice *green* is irrational. Can you explain why?

But if it is irrational for Barbara to choose *green,* then it seems natural for you to assign probability 0 to Barbara wearing *green.* Or, equivalently, to *eliminate* the state *green* from your consideration. Indeed, if you take Barbara seriously as a decision maker, then you must believe that she will choose rationally, and thus assign probability 0 to her making the irrational choice *green.*

From a mechanical perspective, this corresponds to eliminating the column with state *green* in your decision problem in Table 3.1.1, leading to the *reduced* decision problem in Table 3.1.2. Note that in this reduced decision problem, you will always prefer *green* to *red.* In other words, if you assign probability 0 to Barbara wearing the irrational color *green,* that you will definitely prefer *green* to *red.* This makes sense, since we know from the story above that you will always prefer *green* to *red* as long as Barbara does not wear *green* also. In particular, the color *red* can no longer be optimal for you if you believe in Barbara's rationality, that is, believe that Barbara will not make the irrational choice *green.*

But if Barbara anticipates on this type of reasoning by you – and why wouldn't she – and believes that you will choose rationally, then she must believe that you will not choose *red.* Or, more precisely, Barbara will assign probability 0 to you wearing *red.* In that case, in the light of Barbara's conditional preferences in Table 3.1.1, the only optimal choice for Barbara would be *red.*

Thus, if you anticipate on this type of reasoning by Barbara, and believe in Barbara's rationality, then you must assign probability 1 to Barbara wearing *red.* As a consequence, you would definitely choose *blue* yourself.

The reasoning above, which led you to the unique choice *blue,* requires not only that you believe that Barbara chooses rationally, but also that you believe that Barbara believes that you choose rationally, and so on. This type of reasoning is called *common belief in rationality*: As we will see, it will be the central mode of reasoning in this book.

3.2 Belief Hierarchies, Beliefs Diagrams and Types

The purpose of this and the following section will be to offer a formal definition of the central idea of *common belief in rationality.* We will see that its definition naturally involves the idea of *belief hierarchies,* which describe your belief about the opponents' choices, your belief about what the opponents believe about the choices of others, and so on. We start by formally exploring belief hierarchies, and subsequently investigate how such (infinite) belief hierarchies can conveniently be encoded by means of *beliefs diagrams* and epistemic models with *types.*

3.2.1 Belief Hierarchies

Recall the idea of *common belief in rationality.* In a game with two players, *you* and the *opponent,* it means that you believe that the opponent chooses rationally, that you believe that the opponent believes that you choose rationally yourself, and so on. If there are more than two players, then from player i's perspective it means that player i (i) believes that every opponent j chooses rationally, (ii) believes that every opponent j believes that every other player k chooses rationally, and so on.

But what does it mean, precisely, that player i believes that opponent j chooses rationally? It means that i believes that j chooses optimally, given what i believes that j believes about the other players' choices. To express this formally, we need player i's belief about j's choice (a first-order belief), together with i's belief about j's belief about the other players' choices (a second-order belief).

As an illistration, consider again the example "Going to a party" summarized by Table 3.1.1. If you believe that Barbara chooses *yellow* (first-order belief), and believe that Barbara believes that you choose *red* (second-order belief) then you believe that Barbara chooses rationally, because *yellow* is an optimal choice for Barbara if she believes that you choose *red.* Here, when we say "believe", we actually mean "assign probability 1 to",

However, if we would only say that you believe that Barbara chooses *yellow,* then on the basis of this first-order belief alone we cannot conclude whether you believe in Barbara's rationality or not. Indeed, if you were to believe that Barbara believes that you choose *blue,* for instance, then *yellow* would no longer be optimal for Barbara. Hence, to formally express the event that player i believes in j's rationality, we need i's first-order belief about j's choice, together with i's second-order belief about j's belief about the other players' choices.

Similarly, what does it mean that player i believes that opponent j believes that player k chooses rationally? It means that i believes that j believes that k chooses optimally, given what i believes that j believes that k believes about the other players' choices. To express this formally, we thus need i's belief about j's belief about k's choice (a second-order belief), together with i's belief about j's belief about k's belief about the other players' choices (a third-order belief).

By continuing in this fashion, we conclude that the idea of common belief in rationality can only be formalized if we state, for every player i, (i) his belief about the choice of every opponent j (his first-order belief), (ii) his belief about what every opponent j believes about the choice of every other player k (his second-order belief), (iii) his belief about what every opponent j believes about what every opponent k believes about the choice of every other player l (his third-order belief), and so on, *ad infinitum.* Such infinite sequences of beliefs are called *belief hierarchies,* and constitute the language of epistemic game theory. Indeed, throughout this book we will use belief hierarchies to formally define common belief in rationality and other related concepts that are central to epistemic game theory.

Definition 3.2.1 (Belief hierarchy) *A **belief hierarchy** for player i specifies*

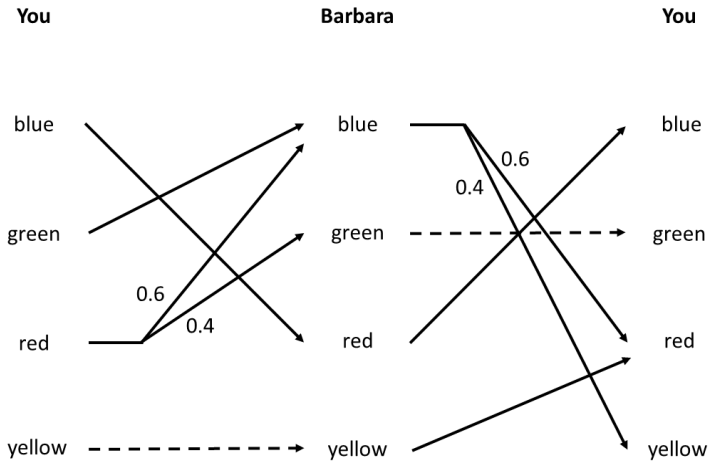

Figure 3.2.1 A beliefs diagram for "Going to a party"

(1) a ***first-order belief,*** *which is a belief about the choices made by i's opponents,*

(2) a ***second-order belief,*** *which is a belief about what every opponent j believes about the choices made by j's opponents,*

(3) a ***third-order belief,*** *which is a belief about what every opponent j believes about what each of his opponents k believes about the choices made by k's opponents,*

and so on.

A practical problem with belief hierarchies, however, is that they contain infinitely many belief levels, which makes it impossible to write them down explicitly. On the other hand, they are indispensable for modelling common belief in rationality, and hence we must find a way to represent them in an easy and compact way. How can this be done? We will see that there are at least two convenient methods for representing belief hierarchies: A graphical representation by means of *beliefs diagrams,* and a mathematical encoding by means of *epistemic models with types.* The first is easy to understand and convenient for examples, whereas the latter is useful for formal definitions and proofs.

3.2.2 Beliefs Diagrams

As an illustration of a beliefs diagram, consider Figure 3.2.1 which has been designed for the example "Going to a party". The arrows represent *beliefs.* For example, the solid arrow from your choice *blue* to Barbara's choice *red* means that you believe that Barbara chooses *red,* and that your choice *blue* is optimal for this belief. The solid forked arrow from your choice *red* to Barbara's choices *blue* and *green,* with probabilities 0.6 and 0.4, represents the probabilistic belief in which you assign probability 0.6 to Barbara choosing *blue* and probability 0.4 to Barbara choosing *green.* Moreover, your choice *red* is optimal for this belief, as we have seen in the previous section. The dashed arrow from your choice *yellow* to Barbara's choice *yellow* means that you believe that Barbara chooses *yellow*, but that your choice *yellow* is *not* optimal for this belief.

Hence, the difference between solid and dashed arrows is that with a solid arrow, the choice from which the arrow starts is *optimal* for the belief represented by the arrow, whereas with a dashed arrow this is not the case. Recall from the first section that your choice *yellow* is not optimal for *any* belief

about Barbara's choice, and hence your choice *yellow* can never be the starting point of a solid arrow. Similarly, the arrows from Barbara's choices to your choices represent beliefs that Barbara has about your choice. Summarizing, the arrows in the beliefs diagram represent *first-order beliefs* about the opponent's choice.

We can also group together two consecutive arrows to obtain *second-order beliefs.* Start, for example, at your choice *blue,* and follow the arrow from your choice *blue* to Barbara's choice *red,* and the succeeding arrow from Barbara's choice *red* to your choice *blue.* These two consecutive arrows together represent the second-order belief in which you believe that Barbara believes that you choose *blue.*

If we start at your choice *red,* then with probability 0.6 we arrive at Barbara's choice *blue,* followed by a forked arrow that points to your choice *red* with probability 0.6 and your choice *yellow* with probability 0.4, and with probability 0.4 it leads to Barbara's choice *green* followed by an arrow to your choice *green.* This induces a more complicated second-order belief, in which you assign probability 0.6 to the event that Barbara believes that you choose *red* and *yellow* with probabilities 0.6 and 0.4, respectively, and in which you assign probability 0.4 to the event that Barbara believes that you choose *green.*

Question 3.2.1 *Describe the second-order belief for you that is obtained if we start at your choice green. Also, describe the second-order belief for Barbara that is obtained if we start at her choice green, and the one that is obtained if we start at her choice blue.*

In a similar fashion, we can also derive *third-order beliefs* from a beliefs diagram by grouping together *three* consecutive arrows. If we start, for instance, at your choice *yellow,* then we first arrive at Barbara's choice *yellow,* with the next arrow leading to your choice *red,* followed by a forked arrow that points to Barbara's choices *blue* and *green* with probabilities 0.6 and 0.4, respectively. The induced third-order belief is that you believe that Barbara believes that you assign probability 0.6 to Barbara choosing *blue* and probability 0.4 to Barbara choosing *green.*

Question 3.2.2 *Describe the third-order belief for you that is obtained if we start at your choice blue. Also, describe the third-order belief for Barbara that is obtained if we start at her choice green.*

Of course we could continue in this way, and derive fourth-order beliefs, fifth-order beliefs, and so on. That is, by starting at an arbitrary choice for you in the beliefs diagram, and following the arrows *ad infinitum,* we can derive a full *belief hierarchy* for you. Similarly for Barbara. In that sense, a beliefs diagram provides a convenient graphical representation of belief hierarchies.

In general, to derive a full belief hierarchy we must be able to follow the arrows *ad infinitum* when we start at a given choice with an outgoing arrow. Therefore, in a beliefs diagram every choice with an outgoing arrow should only point to opponents' choices that have outgoing arrows themselves, so that we never "get stuck" when we follow the arrows. From the beliefs diagram of Figure 3.2.1, for instance, we could create a smaller beliefs diagram by keeping only the arrows from your choice *blue* to Barbara's choice *red,* and the arrow from Barbara's choice *red* to your choice *blue.* Then, the only choices with an outgoing arrow would be your choice *blue* and Barbara's choice *red.* Moreover, by following the arrows when starting at your choice *blue* or Barbara's choice *red* we would never get stuck. From this smaller beliefs diagram, we can therefore derive a full belief hierarchy when we start at your choice *blue* or Barbara's choice *red.*

Question 3.2.3 *Suppose that in the beliefs diagram from Figure 3.2.1 we would delete the arrow from your choice blue to Barbara's choice red, and the arrow from Barbara's choice red to your choice*

You	(b,b)	(g,b)	(r,b)	(y,b)	(b,y)	(g,y)	(r,y)	(y,y)
green	3	0	3	3	3	0	3	3
red	2	2	0	2	2	2	0	2
yellow	1	1	1	0	0	0	0	0

Barbara	(g,b)	(r,b)	(y,b)	(g,y)	(r,y)	(y,y)
blue	0	0	0	3	3	3
green	0	4	4	0	4	4
red	1	0	1	1	0	1
yellow	2	2	0	0	0	0

Chris	(g,b)	(r,b)	(y,b)	(g,g)	(r,g)	(y,g)	(g,r)	(r,r)	(y,r)	(g,y)	(r,y)	(y,y)
blue	0	0	0	2	2	2	2	2	2	2	2	2
yellow	1	1	0	1	1	0	1	1	0	0	0	0

Table 3.2.1 Decision problems in "When Chris joins the party"

blue. Would the remaining collection of arrows qualify as a beliefs diagram? What if we additionally delete the arrow from Barbara's choice green to your choice green?

Beliefs diagrams can also be designed for games with *more than two players*, as the following example will illustrate.

Example 3.2: When Chris joins the party.

It is now one year later, and the same party is organized again. Your friend Chris has heard many nice stories about the party, and would therefore like to join. As before, the problem is what color to wear. Since you have become tired of wearing *blue* all the time, you have decided to give away all your blue clothes to charity. Hence, you can only choose between *green, red* and *yellow* outfits, and your preferences over these three colors have remained the same as before. Barbara, on the other hand, has expanded her wardrobe, but she still only has *blue, green, red* and *yellow* dresses. However, due to new developments in fashion her preferences over these colors have changed. She now prefers *green* to *blue, blue* to *yellow* and *yellow* to *red.* Chris, finally, only has *blue* and *yellow* outfits, and prefers *blue* to *yellow.* Like before, everyone still dislikes wearing the same color as some of the friends. In that case, the persons wearing the same color will be very disappointed.

Suppose that the decision problems for the three players, including their conditional preference relations, are given by the utility matrices in Table 3.2.1. Note that in your decision problem, the states correspond to the choice combinations by Barbara and Chris. For instance, state (g,b) indicates that Barbara chooses *green* and Chris chooses *blue.* Similarly, the states in Barbara's decision problem are the choice combinations by you and Chris. As an example, state (g,b) in Barbara's decision problem specifies that you choose *green* and Chris chooses *blue.* Finally, the states in Chris' decision problem are the choice combinations by you and Barbara, where, for instance, state (g,b) means that you choose *green* and Barbara chooses *blue.*

For this story we can design a beliefs diagram as in Figure 3.2.2. In this diagram, b, g, r and y represent the four colors. As before, the arrows represent beliefs. For instance, the arrow that starts from your choice *green,* leading to the choice pair (b,y) by Barbara and Chris, represents the belief in which you think that Barbara chooses *blue* and Chris chooses *yellow.* Note that your choice *green* is optimal for this belief, and therefore we have drawn a *solid* instead of a *dashed* arrow here.

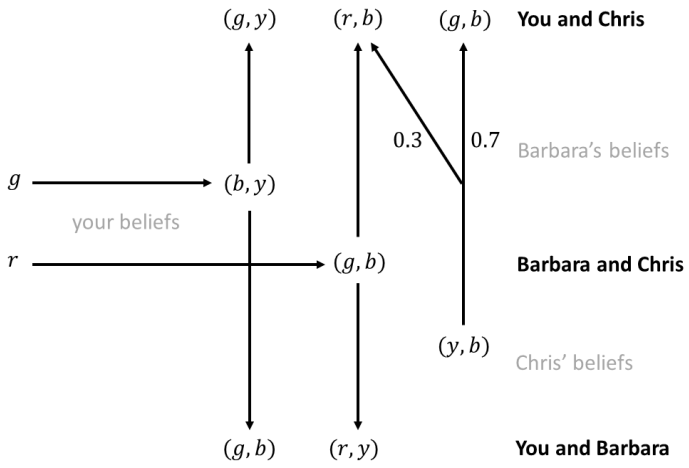

Figure 3.2.2 A beliefs diagram for "When Chris joins the party"

The upwards pointing arrows represent beliefs for Barbara, whereas the downwards pointing arrows represent beliefs for Chris. Consider, for instance, the upwards pointing arrow from Barbara's and Chris' choice combination (b, y) to your and Chris' choice combination (g, y). This represents Barbara's belief in which she thinks that you choose *green* and Chris chooses *yellow.* Note that in the choice combination (b, y) that the arrow starts from, Barbara's choice is *blue.* Since *blue* is optimal for Barbara under the belief just described, the arrow leaving (b, y) is solid. Similarly, the downwards pointing arrow from Barbara's and Chris' choice combination (b, y) to your and Barbara's choice combination (g, b) represents Chris' belief in which he thinks that you choose *green* and Barbara chooses *blue.* As Chris' choice in the starting point (b, y) is *yellow,* and *yellow* is optimal for Chris under the belief just described, also this arrow is solid.

We can also use arrows that represent *probabilistic* beliefs. Consider the forked arrow that starts at Barbara's and Chris' choice combination (y, b), with the associated probabilities 0.3 and 0.7. This arrow represents Barbara's probabilistic belief in which she assigns probability 0.3 to the event that you choose *red* and Chris chooses *blue,* and where she assigns probability 0.7 to the event that you choose *green* and Chris chooses *blue.*

Question 3.2.4 *Explain why Barbara's choice yellow is optimal for the probabilistic belief above.*

The other arrows in the diagram can be interpreted in a similar fashion. Moreover, the reader may verify that for every choice with an outgoing arrow, the choice is optimal for the belief represented by the arrow. This is the reason why we only use *solid* arrows in this beliefs diagram. Note finally that from Barbara's and Chris' choice combination (y, b), there is no downwards pointing arrow representing a belief for Chris. The reason is that Chris' choice *blue* in this choice combination is already present in Barbara's and Chris' choice combination (g, b), which has a downwards pointing arrow. Since the downwards pointing arrow leaving (g, b) serves as an explanation for why Chris could choose *blue,* there is no need to repeat this arrow at Barbara's and Chris' choice combination (y, b).

Similarly as in Figure 3.2.1, we can read the arrows *consecutively* and obtain full *belief hierarchies* for every player in this game. Start, for instance, at your choice *green* and follow the consecutive arrows step by step. The first arrow indicates that you believe that Barbara chooses *blue* and that Chris chooses *yellow.* This is your first-order belief. From Barbara's and Chris' choice combination (b, y) we

can follow Barbara's arrow that points to your and Chris' choice combination (g, y). By putting these two arrows together, we obtain a part of your second-order belief in which you believe that Barbara believes that you choose *green* and that Chris chooses *yellow.* In your third-order belief, what do you believe that Barbara believes that Chris believes that you will do? To answer this question, remember that you believe that Barbara believes that Chris chooses *yellow.* From Chris' arrow leaving Barbara and Chris' choice combination (b, y) we see that Chris chooses *yellow* because he believes that you choose *green* and Barbara chooses *blue.* Hence, you believe that Barbara believes that Chris believes that you choose *green.* By continuing in this fashion, we can also derive the fourth-order beliefs, and all higher order beliefs, if we start at your choice *green* and keep following the consecutive arrows *ad infinitum.* Note that you never get stuck by following the arrows in this way: Every outgoing arrow always points to a choice that has an outgoing arrow itself.

In fact, the above is true no matter where we start in the beliefs diagram. Indeed, if we start at any choice for you, Barbara or Chris, and keep following the consecutive arrows, then we never get stuck. Therefore, we will always be able to derive a full belief hierarchy no matter where we start in the beliefs diagram.

Question 3.2.5 *Suppose we start at your choice red, and keep following the arrows. What is your induced first-order belief about Barbara's choice? In your second-order belief, what do you believe that Chris believes about Barbara's choice? In your third-order belief, what do you believe that Chris believes that Barbara believes about her opponents' choices?*

In the beliefs diagrams from Figures 3.2.1 and 3.2.2, we thus have solid or dashed arrows that go from a choice of player i to choices of the opponent(s). Such arrows may be supplemented by probabilities if they represent probabilistic beliefs. Moreover, to be able to derive *full belief hierarchies,* we must make sure that we never get stuck if we keep following consecutive arrows in the diagram. That is, every arrow must necessarily point to choices that have outgoing (solid or dashed) arrows themselves. These properties give rise to the following general definition of a beliefs diagram.

Definition 3.2.2 (Beliefs diagram) *A beliefs diagram consists of arrows, which point from a choice of some player to choice combinations of his opponents. These arrows represent beliefs about the opponents' choices. If the arrow leaving a certain choice is forked, and reaches several choice combinations by his opponents, then we speak of probabilistic arrows. In that case, the several branches of the arrow must be supplemented with probabilities, and this will then represent a probabilistic belief. Moreover, every arrow must only point to choices that have outgoing arrows themselves. If a choice c_i is optimal for the first-order belief induced by the outgoing arrow, then the outgoing arrow at c_i must be solid. Otherwise, the outgoing arrow at c_i must be dashed.*

In the beliefs diagram of Figure 3.2.1, for instance, the arrow leaving your choice *red* is a probabilistic arrow since it reaches *two* of Barbara's choices, *blue* and *green.* These two choices are supplemented with the probabilities 0.6 and 0.4. Similarly, in the beliefs diagram of Figure 3.2.2, Barbara's arrow leaving Barbara's choice *yellow* is probabilistic as well, since it reaches *two* opponents' choice-combinations, (r, b) and (g, b). The two branches of this arrow are supplemented with the probabilities 0.3 and 0.7. Here, when we mention "Barbara's arrow leaving Barbara's choice *yellow*" we mean Barbara's arrow leaving Barbara's and Chris' choice combination (y, b), in which Barbara chooses *yellow.*

Note also that in a beliefs diagram, not every choice needs to have outgoing arrows, nor does every choice need to be present. In the beliefs diagram of Figure 3.2.2, for instance, your choice *yellow* and Barbara's choice *red* are not present.

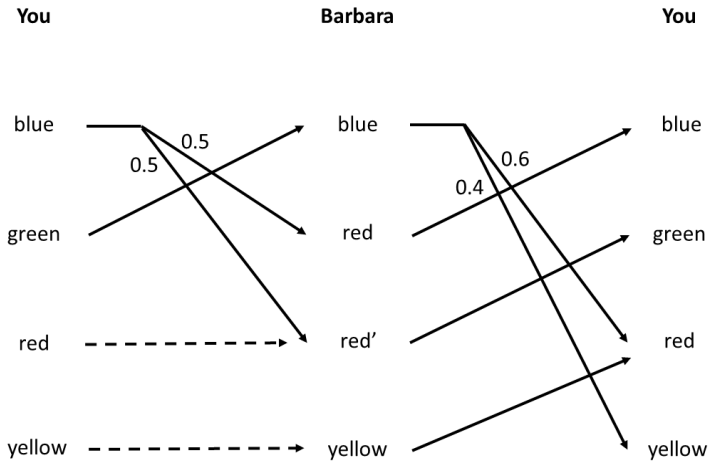

Figure 3.2.3 A beliefs diagram for "Going to a party"

It may also happen that in a beliefs diagram, the same choice of a given player appears more than once. Consider, for instance, the beliefs diagram in Figure 3.2.3 for "Going to a party". There, the choice *red* for Barbara appears twice. The second time it appears, it is denoted by *red'* to distinguish it from the first time it appears.

Let us concentrate on your belief hierarchy that starts at your choice *blue.* With probability 0.5, the arrow goes to Barbara's choice *red,* after which it goes to your choice *blue.* Moreover, with probability 0.5 the arrow leaving your choice *blue* goes to Barbara's choice *red',* after which it goes to your choice *green.* That is, in your belief hierarchy that starts at your choice *blue,* you assign probability 0.5 to the event that Barbara chooses *red* while believing that you choose *blue,* and you assign probability 0.5 to the event that Barbara chooses *red* while believing that you choose *green.* You therefore explain the same choice *red* for Barbara by two different first-order beliefs that Barbara can hold: the belief that you choose *blue,* and the belief that you choose *green.*

In general, if we wish to visualize a belief hierarchy in which you explain the same choice of a player by two, or more, first-order beliefs, we must use a beliefs diagram in which this particular choice appears more than once.

3.2.3 Types

Beliefs diagrams are a very intuitive way to visualize belief hierarchies, and are especially important for illustrative purposes. For rigorous statements and proofs, however, it is desirable to have a more formal encoding of belief hierarchies within a precise mathematical language. This is what we will try to achieve next.

Recall that in a belief hierarchy, player i has a belief about (a) the opponents' choices, (b) the beliefs that the opponents have about *their* opponents' choices, (c) the beliefs that the opponents have about the beliefs that *their* opponents have about the other players' choices, and so on. Note that the beliefs in (b) are the opponents' *first-order* beliefs, that the beliefs in (c) are the opponents' *second-order* beliefs, and so on. Therefore, in a belief hierarchy player i holds a belief about the (a) the opponents' choices, (b) the opponents' first-order beliefs, (c) the opponents' second-order beliefs, and so on. However, the opponents' first-order beliefs, second-order beliefs, and higher-order beliefs

Types	$T_1 = \{t_1^{blue}, t_1^{green}, t_1^{red}, t_1^{yellow}\}$,		$T_2 = \{t_2^{blue}, t_2^{green}, t_2^{red}, t_2^{yellow}\}$
Beliefs for you	$b_1(t_1^{blue})$	=	$(red,\, t_2^{red})$
	$b_1(t_1^{green})$	=	$(blue,\, t_2^{blue})$
	$b_1(t_1^{red})$	=	$(0.6) \cdot (blue,\, t_2^{blue}) + (0.4) \cdot (green,\, t_2^{green})$
	$b_1(t_1^{yellow})$	=	$(yellow,\, t_2^{yellow})$
Beliefs for Barbara	$b_2(t_2^{blue})$	=	$(0.6) \cdot (red,\, t_1^{red}) + (0.4) \cdot (yellow,\, t_1^{yellow})$
	$b_2(t_2^{green})$	=	$(green,\, t_1^{green})$
	$b_2(t_2^{red})$	=	$(blue,\, t_1^{blue})$
	$b_2(t_2^{yellow})$	=	$(red,\, t_1^{red})$

Table 3.2.2 Epistemic model for "Going to a party"

together constitute the opponents' *belief hierarchies.* We thus reach the conclusion that in a *belief hierarchy*, player i holds a probabilistic belief about the opponents' *choices* and the opponents' *belief hierarchies*.

Now, let us call a belief hierarchy a *type.* Then, a *type* for player i holds a belief about the opponents' *choices* and the opponents' *types*. Formally, such a belief takes the form of a *probability distribution* on the opponents' choice-type combinations. That is, a type for player i assigns to every combination of opponents' choices and types some probability, and the sum of all these probabilities must be one. This property leads to the following definition of an *epistemic model* with types.

Definition 3.2.3 (Epistemic model) *An* ***epistemic model*** $M = (T_i, b_i)_{i \in I}$ *specifies*

(a) for every player i a finite set of types T_i, and

(b) for every player i and every type $t_i \in T_i$, a probability distribution $b_i(t_i)$ on the opponents' choice-type combinations. This probability distribution $b_i(t_i)$ represents t_i's belief about the opponents' choices and types.

As an illustration, consider the epistemic model in Table 3.2.2 for the game "Going to a party". As before, you are player 1 and Barbara is player 2. There are four types for you, denoted by t_1^{blue}, t_1^{green}, t_1^{red} and t_1^{yellow}, and four types for Barbara, denoted by t_2^{blue}, t_2^{green}, t_2^{red} and t_2^{yellow}.

The expression $b_1(t_1^{blue}) = (red, t_2^{red})$ indicates that your type t_1^{blue} believes that Barbara chooses *red* and that Barbara has type t_2^{red}. Or, more precisely, your type t_1^{blue} assigns probability 1 to Barbara's choice-type combination $(red,\, t_2^{red})$. The expression

$$b_1(t_1^{red}) = (0.6) \cdot (blue,\ t_2^{blue}) + (0.4) \cdot (green,\ t_2^{green})$$

means that your type t_1^{red} assigns probability 0.6 to Barbara's choice-type combination $(blue,\, t_2^{blue})$, where Barbara chooses *blue* and holds type t_2^{blue}, and assigns probability 0.4 to Barbara's choice-type combination $(green,\, t_2^{green})$, where Barbara chooses *green* and holds type t_2^{green}. In other words, your type t_1^{red} believes that, with probability 0.6, Barbara will choose *blue* while being of type t_2^{blue}, and

that with probability 0.4 Barbara will choose *green* while being of type t_2^{green}. The other beliefs in the epistemic model should be read in a similar way.

Recall that in a beliefs diagram, we could derive a full belief hierarchy for a player if we start at a choice for that player, and keep following the arrows in the diagram. Something similar is true for an epistemic model: If we start at a type for player i, and "keep following the beliefs" in the model, then we will be able to derive the first-order belief, second-order belief, and all higher-order beliefs for that type.

To see how that works, consider your type t_1^{green} in the epistemic model above. Note that type t_1^{green} believes that Barbara chooses *blue* and is of type t_2^{blue}. The induced *first-order* belief, which is the belief about Barbara's choice, is thus that Barbara chooses *blue.* What is the *second-order* belief that your type t_1^{green} has about what Barbara believes about your own choice? Observe that t_1^{green} believes that Barbara is of type t_2^{blue}, and Barbara's type t_2^{blue} assigns probability 0.6 to you choosing *red* and probability 0.4 to you choosing *yellow.* Hence, t_1^{green} believes that Barbara assigns probability 0.6 to you choosing *red* and probability 0.4 to you choosing *yellow.* This is t_1^{green}'s second-order belief.

Writing down t_1^{green}'s *third-order* belief already becomes rather complicated, but is still feasible. As we saw, your type t_1^{green} believes that Barbara is of type t_2^{blue}. In turn, Barbara's type t_2^{blue} assigns probability 0.6 to your type t_1^{red}, which assigns probability 0.6 to Barbara choosing *blue* and probability 0.4 to Barbara choosing *green,* and t_2^{blue} assigns probability 0.4 to your type t_1^{yellow}, which believes that Barbara chooses *yellow.* Summarizing, your type t_1^{green} believes that Barbara assigns probability 0.6 to the event that "you assign probability 0.6 to Barbara choosing *blue* and probability 0.4 to Barbara choosing *green*", and t_1^{green} believes that Barbara assigns probability 0.4 to the event that "you believe that Barbara chooses *yellow*". This is the induced third-order belief. In a similar fashion we can derive all higher-order beliefs, and hence the full belief hierarchy, for type t_1^{green} within the epistemic model.

Question 3.2.6 *What are the induced first-order, second-order and third-order belief for your type t_1^{blue}? And for Barbara's type t_2^{green}?*

For every type in the epistemic model, we can thus derive its full belief hierarchy by "following the consecutive beliefs", much like we derived the belief hierarchies in beliefs diagrams by following the arrows. As such, the epistemic model above – and, in fact, *any* epistemic model – may be viewed as a *formal mathematical encoding* of belief hierarchies.

There is something special about the epistemic model above: It turns out that the belief hierarchies we obtain in this epistemic model are exactly the same as the belief hierarchies from the beliefs diagram in Figure 3.2.1. Indeed, if we start at your type t_1^{green} in the epistemic model, and derive the belief hierarchy by following the beliefs, then we obtain exactly the same belief hierarchy as when we start at your choice *green* in the beliefs diagram, and derive the belief hierarchy by following the arrows. Similarly, if we start at Barbara's choice t_2^{yellow} in the epistemic model and derive the belief hierarchy by following the beliefs, then we obtain the same result as when we start at Barbara's choice *yellow* in the beliefs diagram and keep following the arrows. The same holds for the other types. This also explains the labels we have chosen for the types in the epistemic model.

Since the epistemic model generates exactly the same belief hierarchies as the beliefs diagram from Figure 3.2.1, we may consider the epistemic model above as a *mathematical translation* of this beliefs diagram. Or, put the other way, we may view the beliefs diagram as a *visual translation* of the epistemic model. The underlying reason is that the beliefs about the opponents' choice-type combinations in the epistemic model from Table 3.2.2 are exactly the same beliefs as those represented by the arrows in the beliefs diagram from Figure 3.2.1.

Types	$T_1 = \{t_1^{green}, t_1^{red}\}$, $T_2 = \{t_2^{blue}, t_2^{green}, t_2^{yellow}\}$, $T_3 = \{t_3^{blue}, t_3^{yellow}\}$		
Beliefs for you	$b_1(t_1^{green})$	=	$((blue,\, t_2^{blue}),\, (yellow,\, t_3^{yellow}))$
	$b_1(t_1^{red})$	=	$((green,\, t_2^{green}),\, (blue,\, t_3^{blue}))$
Beliefs for Barbara	$b_2(t_2^{blue})$	=	$((green,\, t_1^{green}),\, (yellow,\, t_3^{yellow}))$
	$b_2(t_2^{green})$	=	$((red,\, t_1^{red}),\, (blue,\, t_3^{blue}))$
	$b_2(t_2^{yellow})$	=	$(0.3) \cdot ((red,\, t_1^{red}),\, (blue,\, t_3^{blue}))+$ $+(0.7) \cdot ((green,\, t_1^{green}),\, (blue,\, t_3^{blue}))$
Beliefs for Chris	$b_3(t_3^{blue})$	=	$((red,\, t_1^{red}),\, (yellow,\, t_2^{yellow}))$
	$b_3(t_3^{yellow})$	=	$((green,\, t_1^{green}),\, (blue,\, t_2^{blue}))$

Table 3.2.3 Epistemic model for "When Chris joins the party"

To see this, consider, for instance, the belief $b_1(t_1^{green}) = (blue,\, t_2^{blue})$ of your type t_1^{green} in the epistemic model. It states that you believe that Barbara chooses *blue* and that Barbara holds the belief hierarchy generated by t_2^{blue}. This, however, is exactly what the arrow leaving your choice *green* in the beliefs diagram tells us. Indeed, this arrow states that you believe that Barbara chooses *blue* and that Barbara holds the belief hierarchy that starts at her choice *blue*. In a similar fashion, it can be verified that *every* belief in the epistemic model of Table 3.2.2 is copied by an arrow in the beliefs diagram in Figure 3.2.1, and *vice versa*. Hence, the epistemic model and the beliefs diagram can be viewed as equivalent representations of the same belief hierarchies.

As an illustration of an epistemic model for three players, consider the epistemic model we designed for "When Chris joins the party", in Table 3.2.3 . You are player 1, Barbara is player 2 and Chris is player 3. There are two types for you, t_1^{green} and t_1^{red}, there are three types for Barbara, t_2^{blue}, t_2^{green} and t_2^{yellow}, and we have two types for Chris, t_3^{blue} and t_3^{yellow}. The belief $b_1(t_1^{green}) = ((blue,\, t_2^{blue}),\, (yellow,\, t_3^{yellow}))$ for your type t_1^{green} indicates that you assign probability 1 to the opponents' choice-type combination $((blue,\, t_2^{blue}),\, (yellow,\, t_3^{yellow}))$ where Barbara chooses *blue,* Barbara has type t_2^{blue}, Chris chooses *yellow,* and Chris has type t_3^{yellow}. In other words, your type t_1^{green} believes that Barbara chooses *blue,* Barbara has type t_2^{blue}, Chris chooses *yellow,* and that Chris has type t_3^{yellow}. The probabilistic belief

$$b_2(t_2^{yellow}) = (0.3) \cdot ((red, t_1^{red}), (blue, t_3^{blue})) + (0.7) \cdot ((green, t_1^{green}), (blue, t_3^{blue}))$$

for Barbara's type t_2^{yellow} assigns probability 0.3 to the choice-type combination $((red, t_1^{red}), (blue, t_3^{blue}))$ where you choose *red,* you have type t_1^{red}, Chris chooses *blue* and Chris has type t_3^{blue}, and assigns probability 0.7 to the choice-type combination $((green, t_1^{green}), (blue, t_3^{blue}))$ where you choose *green,* you have type t_1^{green}, Chris chooses *blue* and Chris has type t_3^{blue}. That is, Barbara's type t_2^{yellow} believes that, with probability 0.3, you choose *red* while having type t_1^{red} and Chris chooses *blue* while having type t_3^{blue}, and t_2^{yellow} believes that, with probability 0.7, you choose *green* while having type t_1^{green} and Chris chooses *blue* while having type t_3^{blue}.

Similarly as in Table 3.2.2, we can derive for every type in the model its first-order belief, second-order belief and all higher-order beliefs, by following the beliefs in the epistemic model. Consider, for instance, your type t_1^{red}, with the belief $b_1(t_1^{red}) = ((green, t_2^{green}), (blue, t_3^{blue}))$. The first-order belief for t_1^{red} is that you believe that Barbara chooses *green* and Chris chooses *blue.* In the induced second-order belief, what do you believe that Barbara believes about Chris' choice? Note that t_1^{red} believes that Barbara holds type t_2^{green}, which believes that Chris chooses *blue.* Hence, your type t_1^{red} believes that Barbara believes that Chris chooses *blue,* which is part of the induced second-order belief. In the induced third-order belief, what do you believe that Chris believes that Barbara believes about your own choice? Observe that t_1^{red} believes that Chris has type t_3^{blue}, which believes that Barbara has type t_2^{yellow}, which in turn assigns probability 0.3 to you choosing *red* and probability 0.7 to you choosing *green.* Therefore, your type t_1^{red} believes that Chris believes that Barbara assigns probability 0.3 to you choosing *red* and probability 0.7 to you choosing *green,* which is part of the induced third-order belief.

Question 3.2.7 *Consider your type t_1^{green}. What is the induced first-order belief about Barbara's and Chris' choice? As part of the second-order belief, what do you believe that Barbara believes about Chris' choice? As part of the third-order belief, what do you believe that Chris believes that you believe about Barbara's choice?*

In the same way, we can derive for every type in the epistemic model its full belief hierarchy by following the beliefs within the model. Note that, similarly as before, also this epistemic model can be viewed as a mathematical translation of a beliefs diagram we saw earlier. Indeed, compare the epistemic model from Table 3.2.3 with the beliefs diagram from Figure 3.2.2. Then, the beliefs of the types in the epistemic model correspond one-to-one to the arrows in the beliefs diagram. Consider, for instance, the belief $b_1(t_1^{green}) = ((blue, t_2^{blue}), (yellow, t_3^{yellow}))$ of your type t_1^{green} in the epistemic model. It states that you believe that Barbara chooses *blue,* that Barbara holds the belief hierarchy induced by t_2^{blue}, that Chris chooses *yellow* and that Chris holds the belief hierarchy induced by t_3^{yellow}. This is exactly what the arrow starting at your choice *green* in the beliefs diagram tells us. Since every belief in the epistemic model has an "equivalent" arrow in the beliefs diagram, and *vice versa,* we may conclude that the epistemic model from Table 3.2.3 and the beliefs diagram from Figure 3.2.2 are equivalent representations of the same belief hierarchies.

All this shows that beliefs diagrams and epistemic models can both be used to encode belief hierarchies in a game. As noted before, beliefs diagrams are especially useful for illustrative purposes, whereas epistemic models are more appropriate for formal statements, definitions and proofs. For that reason, both representations will be extensively used throughout this book.

3.3 Common Belief in Rationality

Remember the idea of common belief in rationality, which states that you believe that every opponent chooses rationally, that you believe that every opponent believes that every other player chooses rationally, and so on. This is a property that pertains to a *belief hierarchy.* Indeed, if you hold a belief hierarchy, then to believe that opponent j chooses rationally means that you believe that j makes a choice which is optimal, given what you think that j believes about his opponents' choices. This condition thus imposes restrictions on your first- and second-order belief. Similarly, to believe that j

believes that another player k chooses rationally imposes restrictions on your second- and third-order belief, and so on.

In the previous section we have seen that belief hierarchies can formally be encoded by means of *types* in an *epistemic model*. In this section we will use this encoding to formally define what common belief in rationality means. We will do so in three steps. First, we define what it means for a choice to be *optimal* for a type. Subsequently, we clarify what it means for a type to *believe in the opponents' rationality.* Finally, we use these building blocks to formally state the idea of *common belief in rationality.*

3.3.1 Optimal Choices for Types

We have seen that every type t_i within an epistemic model holds a probabilistic belief about the opponents' choices and types. In particular, it holds a first-order belief $b_i^1(t_i)$ about the opponents' choices. We then say that a choice c_i is optimal for the type t_i if it is optimal for the first-order belief $b_i^1(t_i)$.

Definition 3.3.1 (Optimal choice for a type) *Consider a type t_i with first-order belief $b_i^1(t_i)$ about the opponents' choice combinations. A choice c_i is **optimal** for a type t_i if*

$$c_i \succsim_{b_i^1(t_i)} c_i'$$

for every choice $c_i' \in C_i$. Or, equivalently, if

$$u_i(c_i, b_i^1(t_i)) \geq u_i(c_i', b_i^1(t_i))$$

for every choice $c_i' \in C_i$.

We often write $u_i(c_i, t_i)$ instead of $u_i(c_i, b_i^1(t_i))$, as an abbreviation. As an illustration, consider the epistemic model in Table 3.2.2 for "Going to a party". Let us focus on your type t_1^{red} for the moment, with the belief

$$b_1(t_1^{red}) = (0.6) \cdot (blue,\ t_2^{blue}) + (0.4) \cdot (green,\ t_2^{green})$$

about Barbara's choice-type combinations. In particular, this type holds the first-order belief

$$b_1^1(t_1^{red}) = (0.6) \cdot blue + (0.4) \cdot green$$

about Barbara's choices. We have seen in Question 3.1.1 that wearing *red* is optimal for this belief. Therefore, choice *red* is optimal for your type t_1^{red}.

Or consider the epistemic model in Table 3.2.3 for "When Chris joins the party". Let us concentrate on Barbara's type t_2^{yellow} with the belief

$$b_2(t_2^{yellow}) = (0.3) \cdot ((red, t_1^{red}), (blue, t_3^{blue})) + (0.7) \cdot ((green, t_1^{green}), (blue, t_3^{blue}))$$

about the opponents' choice-type combinations. This type has the first-order belief

$$b_2^1(t_2^{yellow}) = (0.3) \cdot (red,\ blue) + (0.7) \cdot (green,\ blue)$$

on the opponents' choices, assigning probability 0.3 to the event that you choose *red* and Chris chooses *blue,* and assigning probability 0.7 to the event that you choose *green* and Chris chooses *blue.* We

have seen in Question 3.2.4 that wearing *yellow* is optimal for Barbara under this belief, and therefore choice *yellow* is optimal for Barbara's type t_2^{yellow}.

In the same epistemic model from Table 3.2.3, it may be verified that for you the choices *green* and *red* are optimal for your types t_1^{green} and t_1^{red}, respectively, that for Barbara the choices *blue* and *green* are optimal for her types t_2^{blue} and t_2^{green}, respectively, and that for Chris the choices *blue* and *yellow* are optimal for his types t_3^{blue} and t_3^{yellow}, respectively.

Question 3.3.1 *Consider the epistemic model from Table 3.2.2, designed for "Going to a party". For each of the types in this model, determine the optimal choice(s).*

The notion of an optimal choice for a type will enable us, in the following subsection, to formally define what it means for a type to believe in the opponents' rationality. This will be the central piece in the formal definition of common belief in rationality later.

3.3.2 Belief in the Opponents' Rationality

Intuitively, you believe that opponent j chooses rationally if you believe that j makes a choice that is optimal, given what you think that j thinks that other players will do. How can this condition be formalized within an epistemic model with types?

Consider a type t_i for player i within an epistemic model, which has a probabilistic belief $b_i(t_i)$ about the opponents' choice-type combinations. In particular, $b_i(t_i)$ assigns to every choice-type combination (c_j, t_j) by opponent j some probability. Such a combination (c_j, t_j) may be viewed as an event where player j chooses c_j while holding the belief hierarchy induced by t_j. If type t_i believes in j's rationality, then it must only deem possible events (c_j, t_j) where the choice c_j is optimal for the first-order belief induced by t_j. In other words, $b_i(t_i)$ must only assign positive probability to j's choice-type combinations (c_j, t_j) where the choice c_j is *optimal* for the type t_j. If we require this for every opponent j, we obtain the following definition.

Definition 3.3.2 (Belief in the opponents' rationality) *A type t_i* ***believes in the opponents' rationality*** *if the belief $b_i(t_i)$ on the opponents' choice-type combinations assigns, for every opponent j, only positive probability to choice-type pairs (c_j, t_j) where the choice c_j is optimal for the type t_j.*

Consider, for example, Barbara's type t_2^{yellow} in the epistemic model from Table 3.2.3, with the belief

$$b_2(t_2^{yellow}) = (0.3) \cdot ((red,\ t_1^{red}), (blue,\ t_3^{blue})) + (0.7) \cdot ((green,\ t_1^{green}), (blue,\ t_3^{blue})).$$

This type only assigns positive probability to your choice-type pairs $(red,\ t_1^{red})$ and $(green,\ t_1^{green})$, and to Chris' choice-type pair $(blue,\ t_3^{blue})$. Since we have concluded above that *red* is optimal for your type t_1^{red}, *green* is optimal for your type t_1^{green} and *blue* is optimal for Chris' type t_3^{blue}, we see that Barbara's type t_2^{yellow} believes in her opponents' rationality. In fact, it may be verified that *all* types in this epistemic model believe in the opponents' rationality. Please check this.

This is not true for the epistemic model in Table 3.2.2, however, which has been designed for "Going to a party". Consider, for instance, your type t_1^{red} with the belief

$$b_1(t_1^{red}) = (0.6) \cdot (blue,\ t_2^{blue}) + (0.4) \cdot (green,\ t_2^{green}).$$

In particular, it assigns positive probability to Barbara's choice-type pair $(green,\ t_2^{green})$. Since we have seen in Question 3.3.1 that *green* is not rational for Barbara's type t_2^{green}, we conclude that your type t_1^{red} does not believe in Barbara's rationality.

Question 3.3.2 *Which of the remaining types in the epistemic model from Table 3.2.2 believe in the opponent's rationality? Which do not?*

The definition of belief in the opponents' rationality will now enable us to formally define common belief in rationality.

3.3.3 Common Belief in Rationality

Above we have defined, in a rigorous way, what it means for a type to believe in the opponents' rationality. Having established this definition, it is now surprisingly easy to formalize what it means to believe that opponent j believes in his opponents' rationality. Consider a type t_i for player i in an epistemic model, which has the belief $b_i(t_i)$ on the opponents' choice-type combinations. In particular, $b_i(t_i)$ assigns to every type t_j by opponent j some probability. Such a type t_j may be viewed as the event where player j holds the belief hierarchy induced by t_j. If you believe that j believes in his opponents' rationality, you must only deem possible events t_j where the type t_j believes in his opponents' rationality. In other words, $b_i(t_i)$ must only assign positive probability to j's types t_j where t_j believes in his opponents' rationality. Hence, type t_i believes that his opponents believe in their opponents' rationality if $b_i(t_i)$ only assigns positive probability to opponents' types that believe in their opponents' rationality.

To simplify expressions like this, let us say that type t_i expresses 1-*fold* belief in rationality if t_i believes in the opponents' rationality. Analogously, say that type t_i expresses 2-*fold* belief in rationality if t_i believes that his opponents believe in their opponents' rationality. Within such terminology, we can thus say that type t_i expresses 2-fold belief in rationality if $b_i(t_i)$ only assigns positive probability to opponents' types that express 1-fold belief in rationality.

Similarly, say that type t_i expresses 3-*fold* belief in rationality if t_i believes that every opponent believes that each of his opponents believes in the other players' rationality. Following the pattern above, the formal definition would be that a type t_i expresses 3-fold belief in rationality if $b_i(t_i)$ only assigns positive probability to opponents' types that express 2-fold belief in rationality. By continuing in this way, we can formally define 4-fold belief in rationality, 5-fold belief in rationality, and so on.

Remember that common belief in rationality means that you believe that every opponent chooses rationally, that you believe that every opponent believes that every other player chooses rationally, and so on, *ad infinitum.* Within our terminology above, this would mean that the type t_i expresses 1-fold belief in rationality, 2-fold belief in rationality, 3-fold belief in rationality, and so on, *ad infinitum.* This naturally leads to the following formal definition of common belief in rationality.

Definition 3.3.3 (Common belief in rationality) *A type t_i expresses 1-fold belief in rationality if t_i believes in the opponents' rationality.*

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

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

And so on.

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

As an illustration, consider the epistemic model from Table 3.2.3 for "When Chris joins the party". Consider Barbara's type t_2^{yellow} with the belief

$$b_2(t_2^{yellow}) = (0.3) \cdot ((red,\ t_1^{red}), (blue,\ t_3^{blue})) + (0.7) \cdot ((green,\ t_1^{green}), (blue,\ t_3^{blue}))$$

on the opponents' choice-type combinations. We have already seen that t_2^{yellow} believes in the opponents' rationality, and hence t_2^{yellow} expresses 1-fold belief in rationality. But does t_2^{yellow} also express 2-fold belief in rationality? Note that t_2^{yellow} only assigns positive probability to your types t_1^{red} and t_1^{green}, and to Chris' type t_3^{blue}. Since we have seen that each of these latter types believe in the opponents' rationality – and hence express 1-fold belief in rationality – we conclude that Barbara's type t_2^{yellow} expresses 2-fold belief in rationality.

In fact, it may be verified that *every* type in this epistemic model expresses 2-fold belief in rationality. The reason is very simple. We have seen above that all types in this epistemic model believe in the opponents' rationality, and hence all types in the model express 1-fold belief in rationality. Now take an arbitrary type t_i within the epistemic model. Since $b_i(t_i)$ can only assign positive probability to opponents' types within the model, which all express 1-fold belief in rationality, it automatically follows that t_i expresses 2-fold belief in rationality. This applies to all types, and hence every type within the model will express 2-fold belief in rationality.

But then, all types in the model will also express 3-fold belief in rationality. Indeed, take an arbitrary type t_i within the model. As $b_i(t_i)$ can only assign positive probability to opponents' types within the model which, as we have seen, all express 2-fold belief in rationality, it automatically follows that t_i expresses 3-fold belief in rationality. Again, this applies to all types, and hence every type in the model will express 3-fold belief in rationality.

If we continue in this way, we conclude that all types in the epistemic model from Table 3.2.3 express 4-fold belief in rationality, 5-fold belief in rationality, and so on. Hence, we see that all types in the epistemic model of Table 3.2.3 express *common belief in rationality.* The only thing we used to reach this conclusion is that all types in the model believe in the opponents' rationality.

This logic does not only apply to this particular epistemic model, but carries over to *every* epistemic model for *every* game where all types believe in the opponents' rationality. That is, if we take any game, and any epistemic model for that game where all types believe in the opponents' rationality, then the logic above guarantees that all types in the model will express *common belief in rationality.* Since we will use this property very often throughout this book, we state it as a formal theorem.

Theorem 3.3.1 (Sufficient condition for common belief in rationality) *Consider an epistemic model in which all types believe in the opponents' rationality. Then, all types in the epistemic model will also express common belief in rationality.*

This theorem will often simplify things for us. Indeed, if we design an epistemic model for a given game, and show that every type believes in the opponents' rationality, then the theorem above guarantees that all types will also express common belief in rationality. This is much easier, of course, then explicitly checking for every type that it expresses 1-fold belief in rationality, 2-fold belief in rationality, 3-fold belief in rationality, and so on.

The theorem above also has a counterpart for *beliefs diagrams.* Recall that a beliefs diagram can be used as a visual representation of belief hierarchies. It consists of arrows, which go from a choice c_i of a certain player i to choices of the opponents. Such an arrow is solid if the choice c_i is optimal for the belief represented by the outgoing arrow, whereas it is dashed otherwise.

Suppose now that all arrows in the beliefs diagram are solid. Then, all arrows point to opponents' choices that have solid outgoing arrows themselves. In other words, all arrows point to opponents'

choices that are *optimal* for the beliefs represented by their outgoing arrows. This means, in turn, that in every belief hierarchy generated by the arrows, the corresponding player believes in his opponents' rationality. But then, by the same logic as above, we can conclude that all belief hierarchies generated by the beliefs diagram will express common belief in rationality. Hence, if we start from a beliefs diagram that consists of *solid* arrows only, then all belief hierarchies generated by this diagram will express *common belief in rationality*. This property may be viewed as the counterpart to Theorem 3.3.1 for beliefs diagrams, and is summarized by the theorem below.

Theorem 3.3.2 (Sufficient condition for common belief in rationality) *Consider a beliefs diagram that only contains solid arrows. Then, all belief hierarchies generated by this beliefs diagram express common belief in rationality.*

Consider, for instance, the beliefs diagram in Figure 3.2.2 that has been designed for "When Chris joins the party", and which represents the same belief hierarchies as the epistemic model from Table 3.2.3. Since all arrows in the beliefs diagram are solid, we conclude by the property above that all belief hierarchies generated by the beliefs diagram express common belief in rationality.

Unfortunately, Theorem 3.3.1 above cannot be applied to the epistemic model of Table 3.2.2, designed for "Going to a party". Indeed, we have seen above and in Question 3.3.2 that your type t_1^{red} and Barbara's type t_2^{blue} do not believe in the opponent's rationality, and hence the assumptions in the theorem do not hold. Still we would like to know which types in this epistemic model express common belief in rationality and which do not.

Since the types t_1^{red} and t_2^{blue} do not believe in the opponent's rationality, and hence do not express 1-fold belief in rationality, we can immediately conclude that these two types do not express common belief in rationality. But what about your type t_1^{green}? As t_1^{green} believes that Barbara has type t_2^{blue}, and we have seen that t_2^{blue} does not express 1-fold belief in rationality, we know that t_1^{green} does not express 2-fold belief in rationality, and hence does not express common belief in rationality either.

Question 3.3.3 *Explain, by a similar argument, that also Barbara's type t_2^{yellow} does not express 2-fold belief in rationality.*

Consider next your type t_1^{yellow}, which believes that Barbara is of type t_2^{yellow}. Since we have seen in Question 3.3.3 that Barbara's type t_2^{yellow} does not express 2-fold belief in rationality, your type t_1^{yellow} does not express 3-fold belief in rationality, and hence does not express common belief in rationality either.

Question 3.3.4 *Explain, by a similar argument, that also Barbara's type t_2^{green} does not express 3-fold belief in rationality.*

We thus conclude that your types t_1^{green}, t_1^{red} and t_1^{yellow}, and Barbara's types t_2^{blue}, t_2^{green} and t_2^{yellow}, do not express common belief in rationality.

It remains to explore your type t_1^{blue} and Barbara's type t_2^{red}. Observe that your type t_1^{blue} believes that Barbara is of type t_2^{red}, and that Barbara's type t_2^{red} believes, in turn, that you are of type t_1^{blue}. Hence, a smaller epistemic model only with types t_1^{blue} and t_2^{red} would be sufficient to encode the belief hierarchies of types t_1^{blue} and t_2^{red}. Indeed, consider the smaller epistemic model in Table 3.3.1, where only your type t_1^{blue} and Barbara's type t_2^{red} are present. It can easily be verified that the types t_1^{blue} and t_2^{red} in this smaller epistemic model generate exactly the same belief hierarchies as the types t_1^{blue} and t_2^{red} in the larger epistemic model from Table 3.2.2. The reason is that in the larger model, only the types t_1^{blue} and t_2^{red} – and no other – are needed to derive the belief hierarchies of t_1^{blue} and t_2^{red}.

Types	$T_1 = \{t_1^{blue}\}$,	$T_2 = \{t_2^{red}\}$
Beliefs for you	$b_1(t_1^{blue})$ =	$(red,\, t_2^{red})$
Beliefs for Barbara	$b_2(t_2^{red})$ =	$(blue,\, t_1^{blue})$

Table 3.3.1 Smaller epistemic model for "Going to a party"

Now, in the smaller epistemic model from Table 3.3.1, both types t_1^{blue} and t_2^{red} believe in the opponent's rationality. Please check this. Hence, by Theorem 3.3.1 we can conclude that both types t_1^{blue} and t_2^{red} express common belief in rationality in the epistemic model from Table 3.3.1. Since in the larger epistemic model from Table 3.2.2, the types t_1^{blue} and t_1^{red} generate exactly the same belief hierarchies as in the smaller model from Table 3.3.1, it follows that also in the larger model from Table 3.2.2, both types t_1^{blue} and t_2^{red} express common belief in rationality.

Summarizing, we see that in the epistemic model from Table 3.2.2, designed for "Going to a party", your type t_1^{blue} and Barbara's type t_2^{red} express common belief in rationality, whereas all other types in this model do not.

3.3.4 Rational Choice under Common Belief in Rationality

Intuitively, *common belief in rationality* describes a way of reasoning about the opponents in a game. What we have done in this section so far is to formalize this way of reasoning within a rigorous mathematical framework, relying on epistemic models with types to encode belief hierarchies. However, the central question we wish to answer is: What choices can a player rationally make in a given game if he reasons in accordance with common belief in rationality? In the remainder of this section we will give a precise meaning to this sentence, again using epistemic models with types. In the next section we will then develop an automated elimination procedure that can be used to find the choices a player can rationally make under common belief in rationality, without having to resort to epistemic models.

What does it mean, exactly, when we say that in a given game, player i can rationally make a choice c_i if he reasons in accordance with common belief in rationality? It means that there is some belief hierarchy for player i that expresses common belief in rationality, and such that the choice c_i is optimal for the first-order belief in this belief hierarchy. Since a belief hierarchy can be encoded by a type in some epistemic model, this is equivalent to saying that there is some epistemic model, and some type t_i within that epistemic model, such that the type expresses common belief in rationality, and the choice c_i is optimal for the type t_i.

Definition 3.3.4 (Rational choice under common belief in rationality) *Player i can **rationally make choice** c_i **under common belief in rationality** if there is some epistemic model $M = (T_i, b_i)_{i \in I}$, and some type $t_i \in T_i$ for player i within that model, such that (a) type t_i expresses common belief in rationality, and (b) choice c_i is optimal for the type t_i.*

Similarly, for every $k \geq 1$, we say that player i can *rationally make choice c_i while expressing up to k-fold belief in rationality* if there is some epistemic model $M = (T_i, b_i)_{i \in I}$, and some type $t_i \in T_i$ for player i within that model, such that (a) type t_i expresses up to k-fold belief in rationality, and (b)

Barbara	(g,b)	(r,b)	(g,y)	(r,y)
blue	0	0	3	3
green	0	4	0	4
red	1	0	1	0
yellow	2	2	0	0

Table 3.3.2 Reduced decision problem for Barbara in "When Chris joins the party"

choice c_i is optimal for the type t_i. Here, by expressing up to k-fold belief in rationality we mean that t_i expresses 1-fold belief in rationality, 2-fold belief in rationality, up to and including k-fold belief in rationality.

Let us return to the game "Going to a party", summarized by Table 3.1.1. Which choices can you and Barbara rationally make under common belief in rationality? In Section 3.1 we have argued that under common belief in rationality, it can never be rational for you to wear *green, red* and *yellow,* and it can never be rational for Barbara to wear *blue, green* and *yellow.* In turn, your choice *blue* can be supported by some type that expresses common belief in rationality. Indeed, consider the small epistemic model from Table 3.3.1. We have seen that your type t_1^{blue} expresses common belief in rationality within that model. Since choosing *blue* is optimal for your type t_1^{blue}, we conclude that you can rationally choose *blue* under common belief in rationality. Similarly, Barbara's choice *red* can also be supported by a type that expresses common belief in rationality. In the same epistemic model from Table 3.3.1, Barbara's type t_2^{red} expresses common belief in rationality, as we have seen. As choosing *red* is optimal for her type t_2^{red}, we conclude that also Barbara's choice *red* can rationally be made under common belief in rationality. Altogether, we thus conclude that in the game "Going to a party", your choice *blue* and Barbara's choice *red* are the only choices that can rationally be made under common belief in rationality.

What about the variation of this game, "When Chris joins the party"? What choices can you, Barbara and Chris rationally make under common belief in rationality in this new scenario? A large part of the answer lies at the epistemic model from Table 3.2.3. We have seen that all types in this epistemic model express common belief in rationality. Since for you, choosing *green* is optimal for your type t_1^{green} and choosing *red* is optimal for your type t_1^{red}, you can rationally choose *green* and *red* under common belief in rationality. Similarly, for Barbara choosing *blue* is optimal for her type t_2^{blue}, choosing *green* is optimal for her type t_2^{green} and choosing *yellow* is optimal for her type t_2^{yellow}. Hence, Barbara can rationally choose *blue, green* and *yellow* under common belief in rationality. Finally, for Chris choosing *blue* is optimal for his type t_3^{blue} and choosing *yellow* is optimal for his type t_3^{yellow}, which implies that Chris can rationally choose *blue* and *yellow* under common belief in rationality.

It remains to explore your choice *yellow* and Barbara's choice *red.* These choices are not optimal for any type in the epistemic model from Table 3.2.3. But can these choices perhaps be supported by types in *another* epistemic model that express common belief in rationality?

The answer is "no". To see this, note from your decision problem in Table 3.2.1 that your choice *yellow* is strictly dominated by the randomized choice $(0.4)\cdot$ *green* $+ (0.6)\cdot$ *red.* Please verify this. Hence, we conclude from Theorem 2.7.1 that your choice *yellow* is never optimal for any belief.

Hence, if Barbara believes in your rationality, she will assign probability 0 to your choice *yellow.* Or, equivalently, from Barbara's decision problem in Table 3.2.1 we can eliminate the two states (y,b) and (y,y) where you wear *yellow,* and obtain Barbara's reduced decision problem in Table 3.3.2. In this reduced decision problem, her choice *red* is strictly dominated by the randomized choice $(0.4)\cdot$

blue + (0.6)· *yellow.* Please verify this. Thus, it follows from Theorem 2.7.1 that Barbara's choice *red* cannot be optimal for any belief within this reduced decision problem. In other words, wearing *red* cannot be optimal for Barbara if she believes in your rationality.

Summarizing, we conclude that in the example "When Chris joins the party", you can rationally wear *green* and *red,* but not *yellow,* under common belief in rationality, Barbara can rationally wear *blue, green* and *yellow,* but not *red,* under common belief in rationality, and Chris can rationally wear *blue* and *yellow* under common belief in rationality.

3.4 Recursive Procedure

As stated earlier, an important question we investigate in this chapter is: What choices can a player rationally make in a given game if he reasons in accordance with common belief in rationality? In the previous section we have formally defined what we mean by this sentence. In this, and the following, section we ask: How can we find these choices in an automated way, without having to design an epistemic model with types? Is there some recursive procedure that we can use to easily compute these choices? The answer, as we will see, is "yes". The key to this procedure will be Theorem 2.7.1 from Chapter 2, which characterized those choices that are irrational.

3.4.1 One-Fold Belief in Rationality

We start with a more basic question: What choices can player i rationally make if he expresses 1-fold belief in rationality, that is, believes in the opponents' rationality? In that case, player i will assign probability 0 to the event that his opponents will make an irrational choice. Or, equivalently, player i will eliminate, from his decision problem, all states that involve an irrational choice by an opponent.

Recall from Theorem 2.7.1 that the irrational choices are precisely the choices that are *strictly dominated.* By this, we mean the choices that are either strictly dominated by another choice, or by a randomized choice. Thus, if player i believes in his opponents' rationality, then he will eliminate from his decision problem all states that involve strictly dominated choices by the opponents. We thus obtain a reduced decision problem for player i.

If, subsequently, player i chooses rationally himself then, by the same Theorem 2.7.1, he will make a choice that is not strictly dominated within his reduced decision problem. Hence, we can eliminate from player i's reduced decision problem all choices for player i that are strictly dominated.

By the arguments above, we thus see that the choices that player i can rationally make if he expresses 1-fold belief in rationality are precisely the choices that survive the following procedure: In round 1, we eliminate for all players all choices that are strictly dominated in their decision problems. In round 2, we first eliminate in every decision problem those states that involve strictly dominated choices by the opponents, and subsequently eliminate all choices that are strictly dominated within the resulting reduced decision problem. This procedure is called *two-fold elimination of strictly dominated choices.*

As an illustration, let us apply this procedure to the example "Going to a party" to find the choices that can rationally be made under 1-fold belief in rationality there. The decision problems for you and Barbara can be found in Table 3.1.1, but have been reproduced in Table 3.4.1 for convenience.

We have seen before that for you, *yellow* is strictly dominated by the randomized choice (0.5)· *blue* + (0.5)· *green,* whereas for Barbara the color *green* is strictly dominated by the randomized choice

You	*blue*	*green*	*red*	*yellow*
blue	0	4	4	4
green	3	0	3	3
red	2	2	0	2
yellow	1	1	1	0

Barbara	*blue*	*green*	*red*	*yellow*
blue	0	2	2	2
green	1	0	1	1
red	4	4	0	4
yellow	3	3	3	0

Table 3.4.1 Decision problems for you and Barbara in "Going to a party"

You	*blue*	*green*	*red*	*yellow*
blue	0	4	4	4
green	3	0	3	3
red	2	2	0	2

Barbara	*blue*	*green*	*red*	*yellow*
blue	0	2	2	2
red	4	4	0	4
yellow	3	3	3	0

Table 3.4.2 1-fold reduced decision problems in "Going to a party"

$(0.5)\cdot$ *red* + $(0.5)\cdot$ *yellow.* Thus, in round 1 we can eliminate your choice *yellow* from your decision problem, and Barbara's choice *green* from her decision problem, leading to the 1-fold reduced decision problems in Table 3.4.2.

In round 2, we start by eliminating from your decision problem the state *green,* as it involves an irrational choice by Barbara. In the reduced decision problem that remains, your choice *red* is strictly dominated by *green,* and thus we can eliminate your choice *red* in your decision problem.

Similarly, in Barbara's decision problem we start by eliminating the state *yellow,* as it involves an irrational choice by you. In the reduced decision problem that remains, Barbara's choice *blue* is strictly dominated by her choice *yellow,* and can thus be eliminated.

We finally arrive at the 2-fold reduced decision problems in Table 3.4.3. Thus, under two-fold elimination of strictly dominated choices, the colors *blue* and *green* survive for you, whereas the colors *red* and *yellow* survive for Barbara. As such, under 1-fold belief in rationality, you can rationally choose *blue* and *green,* and Barbara can rationally choose *red* and *yellow.*

3.4.2 Two-Fold Belief in Rationality

We now turn to the next question: What choices can player i rationally make if he expresses 1-fold and 2-fold belief in rationality? That is, if he believes in the opponents' rationality, and believes that the opponents express 1-fold belief in rationality?

We have seen above that if an opponent chooses rationally, and expresses 1-fold belief in rationality, then he will only make choices that survive the 2-fold elimination of strictly dominated choices. Thus, if player i believes in the opponents' rationality, and believes that the opponents express 1-fold belief in rationality, then he will assign probability 0 to all opponents' choices that do *not* survive the 2-fold elimination of strictly dominated choices. Or, equivalently, he will eliminate from his

You	*blue*	*red*	*yellow*
blue	0	4	4
green	3	3	3

Barbara	*blue*	*green*	*red*
red	4	4	0
yellow	3	3	3

Table 3.4.3 Two-fold elimination of strictly dominated choices in "Going to a party"

You	*red*	*yellow*
blue	4	4

Barbara	*blue*	*green*
red	4	4

Table 3.4.4 Three-fold elimination of strictly dominated choices in "Going to a party"

decision problem all states that involve opponents' choices that do not survive the 2-fold elimininination of strictly dominated choices.

This may lead to an even smaller decision problem for player i, with less states than before. If player i then chooses rationally himself, we know by Theorem 2.7.1 that player i will only make choices that are not strictly dominated within this smaller decision problem.

By the arguments above, we conclude that the choices that player i can rationally make if he expresses 1-fold and 2-fold belief in rationality are precisely the choices that survive the following procedure: In rounds 1 and 2, we perform 2-fold elimination of strictly dominated choices. At the beginning of round 3, we eliminate in every decision problem those states that involve opponents' choices that have not survived round 2. In the reduced decision problem so obtained, we then eliminate the choices that are strictly dominated.

Not surprisingly, this procedure is called *three-fold elimination of strictly dominated choices.* By the arguments above, we thus see that the choices that a player can rationally make under 1-fold and 2-fold belief in rationality are precisely the choices that survive the 3-fold elimination of strictly dominated choices.

To see how this procedure works, let us return to the example "Going to a party". We have seen that the 2-fold elimination of strictly dominated choices led to the reduced decision problems in Table 3.4.3. In round 3, we start by eliminating from your decision problem the state *blue,* as Barbara's choice *blue* did not survive round 2. In the smaller decision problem so obtained, your choice *green* is strictly dominated by *blue,* and can thus be eliminated.

Similarly, we can eliminate from Barbara's decision problem the state *red,* as your choice *red* did not survive round 2. In the smaller decision problem so obtained, Barbara's choice *yellow* is strictly dominated by *red,* and can thus be eliminated.

This leads to the reduced decision problems in Table 3.4.4. As you can see, only the color *blue* survives for you, and only the color *red* survives for Barbara. By the arguments above, we may therefore conclude that under 1-fold and 2-fold belief in rationality, you can only rationally choose *blue* and Barbara can only rationally choose *red.*

3.4.3 Common Belief in Rationality

We have seen above that (i) the choices a player can rationally make if he expresses 1-fold belief in rationality are precisely the choices that survive the 2-fold elimination of strictly dominated choices, and (ii) the choices a player can rationally make if he expresses 1-fold and 2-fold belief in rationality are precisely the choices that survive the 3-fold elimination of strictly dominated choices. Of course, we could continue in this fashion, and arrive at the general conclusion that the choices a player can rationally make if he expresses up to k-fold belief in rationality are precisely the choices that survive the $k+1$-fold elimination of strictly dominated choices, for every $k \in \{1, 2, 3, ...\}$.

Here, $k+1$-fold elimination of strictly dominated choices would be the procedure where we first do the k-fold elimination of strictly dominated choices during the first k rounds. Subsequently, at the beginning of round $k+1$ we would eliminate from every decision problem those states that involve

opponents' choices that have not survived round k. This leads to reduced decision problems. Finally, we would at every reduced decision problem eliminate those choices that are strictly dominated.

Since common belief in rationality amounts to k-fold belief in rationality for every k, we conclude, on the basis of these arguments, that the choices a player can rationally make under *common belief in rationality* are precisely the choices that survive all rounds of the *iterated elimination of strictly dominated choices.* By the latter, we mean the procedure where we *keep* eliminating in the way described above until no further states and choices can be eliminated from the decision problems.

This procedure can formally be defined as follows.

Definition 3.4.1 (Iterated elimination of strictly dominated choices) *Start by writing down the decision problems for every player.*

Round 1. *From every decision problem, eliminate those choices that are strictly dominated. This leads to the* 1*-fold reduced decision problems.*

Round 2. *From every* 1*-fold reduced decision problem, eliminate those states that involve opponents' choices that did not survive round 1. Within the (possibly smaller) decision problem so obtained, eliminate all choices that are strictly dominated. This leads to the* 2*-fold reduced decision problems.*

Round 3. *From every* 2*-fold reduced decision problem, eliminate those states that involve opponents' choices that did not survive round 2. Within the (possibly smaller) decision problem so obtained, eliminate all choices that are strictly dominated. This leads to the* 3*-fold reduced decision problems.*

Continue like this until no further states and choices can be eliminated. The choices for a player that eventually remain in his decision problem are said to survive the ***iterated elimination of strictly dominated choices****.*

It should be clear that *k-fold elimination of strictly dominated choices*, for every $k \in \{1, 2, 3, ...\}$, corresponds to the first k rounds of this procedure. Based on our arguments above, we thus arrive at the following characterization of the choices that are possible under common belief in rationality.

Theorem 3.4.1 (Procedure for common belief in rationality) *(a) For every $k \geq 1$, the choices that can rationally be made while expressing up to k-fold belief in rationality are precisely the choices that survive the $k+1$-fold elimination of strictly dominated choices.*

(b) The choices that can rationally be made under common belief in rationality are exactly the choices that survive all rounds of the iterated elimination of strictly dominated choices.

Let us illustrate the procedure and the result above by means of the two examples we have analyzed so far. We first return to "Going to a party". We have already seen that the first three rounds of the procedure lead to the 3-fold reduced decision problems in Table 3.4.4. In round 4, we can still eliminate from your decision problem the state *yellow,* since Barbara's choice *yellow* did not survive round 3. Similarly, from Barbara's decision problem we can still eliminate state *green,* since your choice *green* did not survive round 3 either. We thus arrive at the final decision problems in Table 3.4.5, after which the iterated elimination of strictly dominated choices terminates.

By our arguments above, we thus conclude that under common belief in rationality, you can only rationally wear *blue* whereas Barbara can only rationally wear *red.* This can also be supported by looking at the epistemic model in Table 3.3.1. We have seen that your type t_1^{blue} and Barbara's type t_2^{red} both express common belief in rationality. As your choice *blue* is optimal for your type t_1^{blue}, and Barbara's choice *red* is optimal for her type t_2^{red}, we see that you can indeed rationally choose *blue*

You	*red*
blue	4

Barbara	*blue*
red	4

Table 3.4.5 Iterated elimination of strictly dominated choices in "Going to a party"

You	(b,b)	(g,b)	(r,b)	(y,b)	(b,y)	(g,y)	(r,y)	(y,y)
green	3	0	3	3	3	0	3	3
red	2	2	0	2	2	2	0	2
yellow	1	1	1	0	0	0	0	0

Barbara	(g,b)	(r,b)	(y,b)	(g,y)	(r,y)	(y,y)
blue	0	0	0	3	3	3
green	0	4	4	0	4	4
red	1	0	1	1	0	1
yellow	2	2	0	0	0	0

Chris	(g,b)	(r,b)	(y,b)	(g,g)	(r,g)	(y,g)	(g,r)	(r,r)	(y,r)	(g,y)	(r,y)	(y,y)
blue	0	0	0	2	2	2	2	2	2	2	2	2
yellow	1	1	0	1	1	0	1	1	0	0	0	0

Table 3.4.6 Decision problems in "When Chris joins the party"

under common belief in rationality, and Barbara can rationally choose *red* under common belief in rationality.

To further illustrate the procedure, let us apply it to the game "When Chris joins the party", with the decision problems as stated in Table 3.2.1. For convenience, we have reproduced these decision problems in Table 3.4.6.

Round 1. In your decision problem, we have seen in Section 3.3.4 that your choice *yellow* is strictly dominated by the randomized choice in which you choose *green* with probability 0.4 and *red* with probability 0.6. For Barbara and Chris, no choices are strictly dominated in their respective decision problems. In particular, Barbara's choice *red* is optimal for her belief

$$(0.48) \cdot (g,b) + (0.2) \cdot (y,b) + (0.32) \cdot (g,y)$$

in which she assigns positive probability to three different states. Please check this. Hence, by Theorem 2.7.1, Barbara's choice *red* is not strictly dominated by any randomized choice in her decision problem.

We therefore eliminate your choice *yellow* from your decision problem. This gives rise to the 1-fold reduced decision problems in Table 3.4.7.

Round 2. In Barbara's decision problem, we start by eliminating the states (y,b) and (y,y), that involve your choice *yellow* that did not survive round 1. Subsequently, as we have seen in Section 3.3.4, Barbara's choice *red* is strictly dominated by the randomized choice $(0.4)\cdot$ *blue* $+ (0.6)\cdot$ *yellow*. We can thus eliminate Barbara's choice *red* from her decision problem.

In Chris' decision problem, we can eliminate the states $(y,b), (y,g), (y,r)$ and (y,y) that involve your choice *yellow* which did not survive round 1. But subsequently, no choice for Chris is strictly dominated. We thus obtain the 2-fold reduced decision problems in Table 3.4.8.

You	(b,b)	(g,b)	(r,b)	(y,b)	(b,y)	(g,y)	(r,y)	(y,y)
green	3	0	3	3	3	0	3	3
red	2	2	0	2	2	2	0	2

Barbara	(g,b)	(r,b)	(y,b)	(g,y)	(r,y)	(y,y)
blue	0	0	0	3	3	3
green	0	4	4	0	4	4
red	1	0	1	1	0	1
yellow	2	2	0	0	0	0

Chris	(g,b)	(r,b)	(y,b)	(g,g)	(r,g)	(y,g)	(g,r)	(r,r)	(y,r)	(g,y)	(r,y)	(y,y)
blue	0	0	0	2	2	2	2	2	2	2	2	2
yellow	1	1	0	1	1	0	1	1	0	0	0	0

Table 3.4.7 One-fold reduced decision problems in "When Chris joins the party"

You	(b,b)	(g,b)	(r,b)	(y,b)	(b,y)	(g,y)	(r,y)	(y,y)
green	3	0	3	3	3	0	3	3
red	2	2	0	2	2	2	0	2

Barbara	(g,b)	(r,b)	(g,y)	(r,y)
blue	0	0	3	3
green	0	4	0	4
yellow	2	2	0	0

Chris	(g,b)	(r,b)	(g,g)	(r,g)	(g,r)	(r,r)	(g,y)	(r,y)
blue	0	0	2	2	2	2	2	2
yellow	1	1	1	1	1	1	0	0

Table 3.4.8 Two-fold reduced decision problems in "When Chris joins the party"

You	(b,b)	(g,b)	(y,b)	(b,y)	(g,y)	(y,y)
green	3	0	3	3	0	3
red	2	2	2	2	2	2

Barbara	(g,b)	(r,b)	(g,y)	(r,y)
blue	0	0	3	3
green	0	4	0	4
yellow	2	2	0	0

Chris	(g,b)	(r,b)	(g,g)	(r,g)	(g,y)	(r,y)
blue	0	0	2	2	2	2
yellow	1	1	1	1	0	0

Table 3.4.9 Three–fold reduced decision problems in "When Chris joins the party"

Round 3. In your decision problem, we start by eliminating the states (r,b) and (r,y) that involve Barbara's choice *red* which did not survive round 2. Subsequently, no choice for you is strictly dominated. Similarly, in Chris' decision problem we start by eliminating the states (g,r) and (r,r) that involve Barbara's choice *red* which did not survive round 2. After this, no choice for Chris is strictly dominated. This leads to the 3-fold reduced decision problems in Table 3.4.9.

It can be verified that in each of the 3-fold reduced decision problems, every remaining choice is optimal for some belief. Please check this. Therefore, by Theorem 2.7.1, no remaining choice is strictly dominated, and hence no choices can be eliminated from this point onwards. The procedure thus terminates at round 3, and the surviving choices are *green* and *red* for you, *blue, green* and *yellow* for Barbara, and *blue* and *yellow* for Chris.

We thus know that under common belief in rationality, the only choices that can be chosen rationally are *green* and *red* for you, *blue, green* and *yellow* for Barbara and *blue* and *yellow* for Chris. Recall that, by means of the beliefs diagram in Figure 3.2.2 and the associated epistemic model of Table 3.2.3, we already concluded that these choices can indeed be chosen rationally under common belief in rationality. Therefore, the choices *green* and *red* for you, *blue, green* and *yellow* for Barbara and *blue* and *yellow* for Chris are *precisely* the choices that can rationally be made under common belief in rationality – no more and no less.

3.4.4 Common Belief in Rationality is Always Possible

As we have seen in Theorem 3.4.1, the choices that can rationally be made under common belief in rationality are precisely those that survive the iterated elimination of strictly dominated choices. In particular, for every choice c_i that survives the procedure, we will always be able to construct an epistemic model with a type t_i such that (i) the type t_i expresses common belief in rationality, and (ii) the choice c_i is optimal for the type t_i.

It is easily seen that for every player, there will be at least one choice that survives the iterated elimination of strictly dominated choices. The reason is that at every round of this procedure, it will never happen that *all* remaining choices for a given player are strictly dominated within his decision problem.

To see this, suppose we are in round k of the procedure, and that player i's current decision problem is given by (D_i, D_{-i}, u_i), where D_i is the set of choices for player i that have survived so far, and D_{-i}

the set of opponents' choice combinations that have survived so far. Now, fix an arbitrary belief p_i for player i on D_{-i}, and let c_i be a choice in D_i that is optimal for this belief, among the choices in D_i. Then, by Theorem 2.7.1, this choice c_i will not be strictly dominated in the current decision problem, and hence will not be eliminated at round k.

We thus see that at every round k, and for every player i, there will be at least one choice that survives this round. But since there are only finitely many choices to start with, there must for every player be at least one choice that survives all the elimination rounds. Hence, for every player there is at least one choice that survives the iterated elimination of strictly dominated choices.

This insight, together with Theorem 3.4.1, thus guarantees that for every player i there is a choice c_i, and a type t_i within some epistemic model, such that (i) the type t_i expresses common belief in rationality, and (ii) the choice c_i is optimal for the type t_i. In particular, it will always be able to construct an epistemic model that contains a type t_i which expresses common belief in rationality.

We can actually say a little more: We can always construct an epistemic model where *all* types express common belief in rationality. This is the content of the following theorem.

Theorem 3.4.2 (Common belief in rationality is always possible) *For every game with finitely many choices, we can always find an epistemic model $M = (T_i, b_i)_{i \in I}$ in which all types express common belief in rationality.*

In the following subsection we will show, by means of an example, how the procedure can be used to construct an epistemic model where all types express common belief in rationality.

3.4.5 Using the Procedure to Construct Epistemic Models

Suppose we apply the *iterated elimination of strictly dominated choices* to a game. This results in a final decision problem for each of the players. By construction of the procedure, every choice in a final decision problem will not be strictly dominated. Hence, by Theorem 2.7.1, every surviving choice will be optimal for some belief over the remaining states in the final decision problem.

Moreover, the states in the final decision problems will only involve opponents' choices that are not strictly dominated themselves. In other words, these states only contain choices for an opponent that are optimal for some belief within that opponent's final decision problem. And so on.

The situation above can be expressed by a beliefs diagram in which (i) the choices that appear are precisely the choices that survive the procedure, and (ii) all arrows are solid. This beliefs diagram, in turn, can be translated into an epistemic model where all types believe in the opponents' rationality. By Theorem 3.3.1 we thus know that all types in this epistemic model will express common belief in rationality.

To see how this works in practice, let us return to the example "When Chris joins the party". In Section 3.4.3 we applied the *iterated elimination of strictly dominated choices* to this game, and saw that the final decision problems were given by Table 3.4.9.

Within your final decision problem, your choice *green* is optimal for the belief (*blue, yellow*) whereas your choice *red* is optimal for the belief (*green, blue*).

Within Barbara's final decision problem, her choice *blue* is optimal for the belief (*green, yellow*), her choice *green* is optimal for the belief (*red, blue*), and her choice *yellow* is optimal for the belief $(0.3)\cdot$ (*red, blue*) + $(0.7)\cdot$ (*green, blue*).

Within Chris' final decision problem, his choice *blue* is optimal for the belief (*red, yellow*), whereas his choice *yellow* is optimal for the belief (*green, blue*).

These beliefs give rise to the beliefs diagram in Figure 3.2.2, which contains only solid arrows. This beliefs diagram, in turn, can be translated into the epistemic model of Table 3.2.3, where all types express common belief in rationality.

In this way we can always construct, for every game, (i) a beliefs diagram that involves precisely those choices that survive the iterated elimination of strictly dominated choices, and where every choice has a solid outgoing arrow, and (ii) an epistemic model where all types express common belief in rationality, and that contains, for every choice surviving the procedure, a type for which that choice is optimal.

3.5 Order of Elimination

By definition, the procedure of *iterated elimination of strictly dominated choices* requires us to eliminate, at *every* round and at *every* decision problem, *all* states that involve opponents' choices which did not survive the previous round, and, subsequently, *all* choices that are strictly dominated for that player in his current decision problem. Suppose now that at some of the rounds, and at some of the decision problems, we do not eliminate *all* states and choices that we could. Does it matter for the final output of the procedure? We will see that the answer is "no".

Before we explain why, let us first illustrate this issue by the example "Going to a party" with the decision problems as given in Table 3.1.1. If we apply the *iterated elimination of strictly dominated choices* at full speed – that is, always eliminate *all* states and choices that we can – then the following eliminations will occur:

In round 1, we eliminate your choice *yellow* from your decision problem because it is strictly dominated by a randomized choice that assigns probability 0.5 to *blue* and *green,* and for a similar reason we eliminate Barbara's choice *green* from her decision problem.

In round 2, we first eliminate state *green* from your decision problem, and subsequently eliminate your choice *red* there because it has become strictly dominated by *green.* Similarly, in Barbara's decision problem we first eliminate the state *yellow,* after which we eliminate Barbara's choice *blue* there because it has become strictly dominated by her choice *yellow.*

In round 3, we first eliminate the state *blue* from your decision problem, and subsequently we eliminate your choice *green* there because it is strictly dominated by *blue.* In Barbara's decision problem, we first eliminate the state *red,* after which we eliminate Barbara's choice *yellow* because it is strictly dominated by her choice *red.* After this round, no more states and choices can be eliminated from the decision problems.

Suppose now that at round 1, we only eliminate your strictly dominated choice *yellow,* but *not* Barbara's strictly dominated choice *green.* Then, at round 2, the state *green* will still be present in your decision problem. Therefore, we cannot eliminate your choice *red* in round 2, because it is not strictly dominated in your current decision problem. In fact, none of your choices in the current decision problem are strictly dominated.

Assume that from this moment on, we always eliminate everything we can. This gives rise to the reduced decision problems in Table 3.5.1. We thus see that with this alternative order of elimination, in which we did not eliminate Barbara's strictly dominated choice *green* at the first round, only the choices *blue* for you and *red* for Barbara survive – exactly the same choices that survived under the full speed elimination order above. However, with the alternative order of elimination it took five rounds, instead of four, until the procedure terminated. But the eventual output is exactly the same.

Start

You	*blue*	*green*	*red*	*yellow*
blue	0	4	4	4
green	3	0	3	3
red	2	2	0	2
yellow	1	1	1	0

Barbara	*blue*	*green*	*red*	*yellow*
blue	0	2	2	2
green	1	0	1	1
red	4	4	0	4
yellow	3	3	3	0

Round 1

You	*blue*	*green*	*red*	*yellow*
blue	0	4	4	4
green	3	0	3	3
red	2	2	0	2

Barbara	*blue*	*green*	*red*	*yellow*
blue	0	2	2	2
green	1	0	1	1
red	4	4	0	4
yellow	3	3	3	0

Round 2

You	*blue*	*green*	*red*	*yellow*
blue	0	4	4	4
green	3	0	3	3
red	2	2	0	2

Barbara	*blue*	*green*	*red*
red	4	4	0
yellow	3	3	3

Round 3

You	*red*	*yellow*
blue	4	4

Barbara	*blue*	*green*	*red*
red	4	4	0
yellow	3	3	3

Round 4

You	*red*	*yellow*
blue	4	4

Barbara	*blue*
red	4

Round 5

You	*red*
blue	4

Barbara	*blue*
red	4

Table 3.5.1 Changing the order of elimination in "Going to a party"

Question 3.5.1 *Describe an alternative order of elimination in "Going to a party" in which at every round you only eliminate one strictly dominated choice in total. What choices survive this procedure? How many rounds does it take for the procedure to terminate?*

We will now see that this is not a coincidence: No matter which game we take, and no matter which order of elimination we choose for the *iterated elimination of strictly dominated choices,* the surviving choices will always be the same as under the full speed elimination order.

Theorem 3.5.1 (Order independence) *Changing the order of elimination in the iterated elimination of strictly dominated choices does not change the sets of choices that survive the procedure at the end.*

This result has an important practical implication: If, for a given game, we apply the procedure, and at some steps forget to eliminate some states or choices, then eventually we are still guaranteed to end up with the correct result. Provided, of course, we do not forget to eliminate certain choices or states *forever.*

However, part (a) of Theorem 3.4.1 may no longer hold if we replace the full speed elimination procedure by an alterative order of elimination. Part (a) states that the choices that can rationally be made while expressing up to k-fold belief in rationality are precisely the choices that survive the first $k+1$ rounds of the *iterated elimination of strictly dominated choices.* This only holds for the *full speed* elimination procedure, however. To see why, consider the game "Going to a party" with the alternative elimination procedure presented at the beginning of this section. Note that your choice *red* survives the first two rounds of this alternative procedure. However, you cannot rationally choose *red* under 1-fold belief in rationality. Indeed, if you believe in Barbara's rationality, you must believe that Barbara will not choose *green,* and hence your choice *green* will be better than your choice *red.*

3.6 Proofs

3.6.1 Proofs of Section 3.3

Proof of Theorem 3.3.1. Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ in which all types believe in the opponents' rationality. We show, by induction on k, that all types express k-fold belief in rationality for all $k \geq 1$.

By definition, all types express 1-fold belief in rationality because they believe in the opponents' rationality.

Take some $k \geq 2$ and assume that all types express $(k-1)$-fold belief in rationality. Since a type t_i can only assign positive probability to other types in the model M, it follows by the induction assumption that every type t_i only assigns positive probability to opponents' types that express $(k-1)$-fold belief in rationality. Hence, every type t_i in the model expresses k-fold belief in rationality. By induction on k, we conclude that all types express k-fold belief in rationality for every k. That is, all types in M express common belief in rationality. ■

Proof of Theorem 3.3.2. Consider a beliefs diagram that only contains solid arrows. We show, by induction on k, that all belief hierarchies generated by this beliefs diagram express k-fold belief in rationality for all $k \geq 1$.

For $k = 1$, consider the belief hierarchy that starts at some choice c_i in the beliefs diagram. Since the beliefs diagram only contains solid arrows, every arrow leaving c_i points to an opponent's choice c_j with solid outgoing arrows. That is, the belief hierarchy starting at c_i only assigns positive probability to combinations of choices c_j and first-order beliefs for j where the choice c_j is optimal for the first-order belief. Hence, the belief hierarchy starting at c_i expresses 1-fold belief in rationality.

Suppose now that $k \geq 2$ and that every belief hierarchy generated within this beliefs diagram expresses $(k-1)$-fold belief in rationality. Consider the belief hierarchy that starts at some choice c_i in the beliefs diagram. Then, every arrow leaving c_i points to some choice c_j for some opponent j. Since, by the induction assumption, the belief hierarchy starting at c_j expresses $(k-1)$-fold belief in rationality, it follows that the belief hierarchy starting at c_i only assigns positive probability to opponents' belief hierarchies that express $(k-1)$-fold belief in rationality. Hence, the belief hierarchy starting at c_i expresses k-fold belief in rationality. This holds for every belief hierarchy generated by the beliefs diagram. Hence, it follows by induction on k that all belief hierarchies generated by the beliefs diagram express common belief in rationality. ■

3.6.2 Proofs of Section 3.4

To prove Theorem 3.4.1 we need a preparatory result. For every round k and every player i, let C_{-i}^k be the set of states that survive the first k rounds of the *iterated elimination of strictly dominated choices* in player i's decision problem. Moreover, let C_i^k denote the set of choices for player i that survive the first k rounds of the procedure in i's decision problem. Hence, if $k \geq 1$, every choice $c_i \in C_i^k$ belongs to C_i^{k-1} and is not strictly dominated in the decision problem $(C_i^{k-1}, C_{-i}^k, u_i)$. By Theorem 2.7.1, we thus know that every choice $c_i \in C_i^k$ is optimal for some belief b_i in $(C_i^{k-1}, C_{-i}^k, u_i)$. That is,

$$u_i(c_i, b_i) \geq u_i(c_i', b_i) \text{ for all } c_i' \in C_i^{k-1}.$$

We will show, however, that the inequality above holds for every $c_i' \in C_i$, and not only for every $c_i' \in C_i^{k-1}$. Hence, we will prove that for every choice $c_i \in C_i^k$ there is a belief b_i in $\Delta(C_{-i}^k)$ such that

$$u_i(c_i, b_i) \geq u_i(c_i', b_i) \text{ for all } c_i' \in C_i.$$

Or, equivalently, every choice $c_i \in C_i^k$ is optimal for some belief b_i in the larger decision problem (C_i, C_{-i}^k, u_i). This is the content of the following lemma.

Lemma 3.6.1 (Optimality property) *For every player i and every round $k \geq 0$, let C_i^k be the set of choices for player i that survive the first k rounds of the iterated elimination of strictly dominated choices, and let C_i^* be the set of choices that survive all rounds, in player i's decision problem. Similarly, let C_{-i}^k be the set of states that survive the first k rounds, and let C_{-i}^* be the set of opponents' choice-combinations that survive all rounds, in player i's decision problem.*

(a) For every $k \geq 1$, a choice c_i is in C_i^k if and only if c_i is optimal for some belief in (C_i, C_{-i}^k, u_i).

(b) A choice c_i is in C_i^* if and only if c_i is optimal for some belief in (C_i, C_{-i}^*, u_i).

Proof of Lemma 3.6.1. (a) We prove the statement by induction on k. By definition, C_i^1 contains precisely those choices in C_i that are not strictly dominated in $(C_i, C_{-i}, u_i) = (C_i, C_{-i}^0, u_i)$. Hence, by Theorem 2.7.1, $c_i \in C_i^1$ exactly when c_i is optimal for some belief in (C_i, C_{-i}^0, u_i). Therefore, the statement is true for $k = 1$.

Suppose now that $k \geq 2$ and that the statement is true for $k-1$. To show the "only if" direction for k, consider some choice $c_i \in C_i^k$. Then, by definition, c_i is in C_i^{k-1} and is not strictly dominated within $(C_i^{k-1}, C_{-i}^k, u_i)$. By Theorem 2.7.1, c_i is optimal for some belief b_i within the decision problem $(C_i^{k-1}, C_{-i}^k, u_i)$. Hence,

$$u_i(c_i, b_i) \geq u_i(c_i', b_i) \text{ for all } c_i' \in C_i^{k-1}. \tag{3.6.1}$$

Let $c_i^* \in C_i$ be optimal for the belief b_i within (C_i, C_i^k, u_i). That is,

$$u_i(c_i^*, b_i) \geq u_i(c_i', b_i) \text{ for all } c_i' \in C_i. \tag{3.6.2}$$

As $b_i \in \Delta(C_{-i}^k)$ and $C_{-i}^k \subseteq C_{-i}^{k-1}$, it follows that $b_i \in \Delta(C_{-i}^{k-1})$. Hence, c_i^* is optimal for the belief b_i within (C_i, C_{-i}^{k-1}, u_i). But then, we know by the induction assumption that that $c_i^* \in C_i^{k-1}$. By (3.6.1) we thus conclude that

$$u_i(c_i, b_i) \geq u_i(c_i^*, b_i). \tag{3.6.3}$$

By combining (3.6.3) and (3.6.2) we see that

$$u_i(c_i, b_i) \geq u_i(c_i^*, b_i) \geq u_i(c_i', b_i) \text{ for all } c_i' \in C_i,$$

and hence c_i is optimal for the belief b_i in (C_i, C_{-i}^k, u_i). We thus have shown that every $c_i \in C_i^k$ is optimal for some belief in (C_i, C_{-i}^k, u_i). This establishes the "only if" part.

To show the "if" part, consider some choice c_i that is optimal for some belief b_i in (C_i, C_{-i}^k, u_i). Since $b_i \in \Delta(C_{-i}^k)$ and $C_{-i}^k \subseteq C_{-i}^{k-1}$, it follows that $b_i \in \Delta(C_{-i}^{k-1})$. Hence, c_i is optimal for the belief b_i within (C_i, C_{-i}^{k-1}, u_i). By the induction assumption we conclude that $c_i \in C_i^{k-1}$. Since c_i is optimal for b_i in (C_i, C_{-i}^k, u_i) and $c_i \in C_i^{k-1}$, it follows in particular that c_i is optimal for b_i in $(C_i^{k-1}, C_{-i}^k, u_i)$. Hence, by Theorem 2.7.1, c_i is not strictly dominated within $(C_i^{k-1}, C_{-i}^k, u_i)$. This means, in turn, that $c_i \in C_i^k$. We thereby have shown that every choice c_i that is optimal for some belief in (C_i, C_{-i}^k, u_i), must be in C_i^k. This establishes the "if" 'direction.

By combining the "only if" and "if" direction, the statement in (a) follows for k. By induction on k, statement (a) holds for every $k \geq 1$.

(b) Suppose that the procedure terminates at the end of round K. That is, $C_i^* = C_i^K = C_i^{K+1}$ and $C_{-i}^* = C_{-i}^K = C_{-i}^{K+1}$ for every player i. Then, c_i is in C_i^* precisely when $c_i \in C_i^{K+1}$. By applying (a) to

$k = K+1$, we know that c_i is in C_i^{K+1} precisely when c_i is optimal for some belief in $(C_i, C_{-i}^{K+1}, u_i) = (C_i, C_{-i}^*, u_i)$. Hence, c_i is in C_i^* if and only if c_i is optimal for some belief in (C_i, C_{-i}^*, u_i). This completes the proof. ■

We can now use Lemma 3.6.1 to prove Theorem 3.4.1.

Proof of Theorem 3.4.1. As an additional piece of notation, let BR_i^k denote the set of choices that player i can rationally make while expressing up to k-fold belief in rationality. Hence, to prove (a) in Theorem 3.4.1 we must show that $BR_i^k = C_i^{k+1}$ for every player i and every $k \geq 1$. We show this in two steps: (i) prove that $BR_i^k \subseteq C_i^{k+1}$ for all $k \geq 1$, and (ii) prove that $C_i^{k+1} \subseteq BR_i^k$ for all $k \geq 1$.

(i) Show that $BR_i^k \subseteq C_i^{k+1}$ for all players i and all $k \geq 1$.

We prove this by induction on k. For $k = 1$, take some $c_i \in BR_i^1$. Then, there is some epistemic model $M = (T_i, b_i)_{i \in I}$ and some type $t_i \in T_i$ such that t_i expresses 1-fold belief in rationality and c_i is optimal for t_i. Suppose that $b_i(t_i)$ assigns positive probability to some opponent's choice-type pair (c_j, t_j). Since t_i expresses 1-fold belief in rationality, c_j must be optimal for t_j. Hence, c_j is optimal for t_j's first-order belief in the full decision problem (C_j, C_{-j}, u_j) which, by Lemma 3.6.1, implies that $c_j \in C_j^1$. Hence, t_i's first-order belief only assigns positive probability to opponents' choices c_j which are in C_j^1, and thus only assigns positive probability to states in C_{-i}^1. As c_i is optimal for t_i, we conclude that c_i is optimal for t_i's first-order belief in (C_i, C_{-i}^1, u_i) which implies, by Lemma 3.6.1, that c_i is in C_i^2. We thus have shown that every choice $c_i \in BR_i^1$ must be in C_i^2, and hence $BR_i^1 \subseteq C_i^2$ for all players i.

Now suppose that $k \geq 2$ and that, by the induction assumption, $BR_i^{k-1} \subseteq C_i^k$ for all players i. Consider some player i and some $c_i \in BR_i^k$. Then, there is some epistemic model $M = (T_i, b_i)_{i \in I}$ and some type $t_i \in T_i$ such that t_i expresses up to k-fold belief in rationality and c_i is optimal for t_i. Suppose that $b_i(t_i)$ assigns positive probability to some opponent's choice-type pair (c_j, t_j). Since t_i expresses up to k-fold belief in rationality, the choice c_j must be optimal for t_j and t_j must express up to $(k-1)$-fold belief in rationality. Hence, $c_j \in BR_j^{k-1}$. Since, by the induction assumption, $BR_j^{k-1} \subseteq C_j^k$, we know that $c_j \in C_j^k$. We thus conclude that t_i's first-order belief only assigns positive probability to opponents' choices c_j that are in C_j^k, and hence only assigns positive probability to states in C_{-i}^k. As c_i is optimal for t_i, we conclude that c_i is optimal for t_i's first-order belief in (C_i, C_{-i}^k, u_i), which implies, by Lemma 3.6.1, that c_i is in C_i^{k+1}. We thus have shown that every choice $c_i \in BR_i^k$ must be in C_i^{k+1}, and hence $BR_i^k \subseteq C_i^{k+1}$. By induction on k, we conclude that $BR_i^k \subseteq C_i^{k+1}$ for all players i and all $k \geq 1$. This completes the proof of (i).

(ii) Show that $C_i^{k+1} \subseteq BR_i^k$ for all players i and all $k \geq 1$.

Hence, for every choice $c_i \in C_i^{k+1}$ we must show that there is some epistemic model, and some type $t_i^{c_i}$ in it, such that $t_i^{c_i}$ expresses up to k-fold belief in rationality and c_i is optimal for $t_i^{c_i}$. We will now construct a *single* epistemic model $M = (T_i, b_i)_{i \in I}$ that contains *all* such types. For every player i, define the set of types

$$T_i = \{t_i^{c_i} \mid c_i \in C_i^1\}.$$

That is, for every choice c_i that survives at least one round of the procedure, we define a type $t_i^{c_i}$. To define the beliefs of these types about the opponents' choice-type combinations we distinguish the following three cases, assuming that the procedure terminates at the end of round K.

Case 1. Suppose that $c_i \in C_i^1 \backslash C_i^2$. Then, by Lemma 3.6.1, c_i is optimal for some belief $b_i^{c_i} \in \Delta(C_{-i})$ within (C_i, C_{-i}, u_i). For every opponent j choose some arbitrary type $\hat{t}_j \in T_j$, and define

$$b_i(t_i^{c_i})((c_j, t_j)_{j \neq i}) := \begin{cases} b_i^{c_i}((c_j)_{j \neq i}), & \text{if } t_j = \hat{t}_j \text{ for all } j \neq i \\ 0, & \text{otherwise} \end{cases} \tag{3.6.4}$$

for all $(c_j, t_j)_{j \neq i}$ in $C_{-i} \times T_{-i}$.

Case 2. Suppose that $c_i \in C_i^k \backslash C_i^{k+1}$ for some $k \in \{2, ..., K-1\}$. Then, by Lemma 3.6.1, c_i is optimal for some belief $b_i^{c_i} \in \Delta(C_{-i}^{k-1})$ within (C_i, C_{-i}^{k-1}, u_i). Define

$$b_i(t_i^{c_i})((c_j, t_j)_{j \neq i}) := \begin{cases} b_i^{c_i}((c_j)_{j \neq i}), & \text{if } c_j \in C_j^{k-1} \text{ and } t_j = t_j^{c_j} \text{ for all } j \neq i \\ 0, & \text{otherwise} \end{cases} \tag{3.6.5}$$

for all $(c_j, t_j)_{j \neq i}$ in $C_{-i} \times T_{-i}$.

Case 3. Suppose that $c_i \in C_i^K$. As the procedure terminates at round K we have that $c_i \in C_i^*$. Hence, by Lemma 3.6.1, c_i is optimal for some belief $b_i^{c_i} \in \Delta(C_{-i}^*)$ within (C_i, C_{-i}^*, u_i). Define

$$b_i(t_i^{c_i})((c_j, t_j)_{j \neq i}) := \begin{cases} b_i^{c_i}((c_j)_{j \neq i}), & \text{if } c_j \in C_j^* \text{ and } t_j = t_j^{c_j} \text{ for all } j \neq i \\ 0, & \text{otherwise} \end{cases} \tag{3.6.6}$$

for all $(c_j, t_j)_{j \neq i}$ in $C_{-i} \times T_{-i}$. This completes the construction of the epistemic model $M = (T_i, b_i)_{i \in I}$.

Note that in this epistemic model, every type $t_i^{c_i}$ holds the first-order belief $b_i^{c_i}$. As, by definition, c_i is optimal for $b_i^{c_i}$ within (C_i, C_{-i}, u_i), we conclude that c_i is optimal for $t_i^{c_i}$, for every player i and every $c_i \in C_i^1$.

We now show that for every $k \geq 2$ and every choice $c_i \in C_i^k$, the associated type $t_i^{c_i}$ expresses up to $(k-1)$-fold belief in rationality. We show this by induction on k.

For $k = 2$, consider some choice $c_i \in C_i^2$ and the associated type $t_i^{c_i}$ with the belief given by (3.6.5) or (3.6.6). By (3.6.5) and (3.6.6), the belief $b_i(t_i^{c_i})$ only assigns positive probability to opponent's choice-type pairs $(c_j, t_j^{c_j})$ where $c_j \in C_j^1$. As c_j is optimal for $t_j^{c_j}$, the type $t_i^{c_i}$ only assigns positive probability to opponent's choice-type pairs $(c_j, t_j^{c_j})$ where c_j is optimal for $t_j^{c_j}$. Hence, $t_i^{c_i}$ expresses 1-fold belief in rationality. This holds for every type $t_i^{c_i}$ where $c_i \in C_i^2$.

Suppose now that $k \geq 3$ and that, by the induction assumption, $t_i^{c_i}$ expresses up to $(k-2)$-fold belief in rationality for every $c_i \in C_i^{k-1}$ and every player i. Consider some choice $c_i \in C_i^k$ and the associated type $t_i^{c_i}$ with the belief given by (3.6.5) or (3.6.6). By (3.6.5) and (3.6.6) it follows that $b_i(t_i^{c_i})$ only assigns positive probability to opponent's choice-type pairs $(c_j, t_j^{c_j})$ where $c_j \in C_j^{k-1}$. By the induction assumption we know that $t_j^{c_j}$ expresses up to $(k-2)$-fold belief in rationality. As c_j is optimal for $t_j^{c_j}$, we conclude that $t_i^{c_i}$ only assigns positive probability to opponent's choice-type pairs $(c_j, t_j^{c_j})$ where c_j is optimal for $t_j^{c_j}$, and $t_j^{c_j}$ expresses up to $(k-2)$-fold belief in rationality. Hence, $t_i^{c_i}$ expresses up to $(k-1)$-fold belief in rationality. This holds for every type $t_i^{c_i}$ where $c_i \in C_i^k$.

By induction on k, we conclude that for every $k \geq 2$ and every choice $c_i \in C_i^k$, the associated type $t_i^{c_i}$ expresses up to $(k-1)$-fold belief in rationality.

We next show that for every $c_i \in C_i^K$, the associated type $t_i^{c_i}$ expresses *common* belief in rationality. Consider the smaller epistemic model $M^* = (T_i^*, b_i)_{i \in I}$ where the set of types for player i is

$$T_i^* := \{t_i^{c_i} \mid c_i \in C_i^*\},$$

and the beliefs of the types are given by (3.6.6). Note that this is a well-defined epistemic model, since by (3.6.6) every type $t_i^{c_i} \in T_i^*$ with $c_i \in C_i^*$ only assigns positive probability to opponent's types $t_j^{c_j} \in T_j^*$ where $c_j \in C_j^*$. We show that every type in M^* believes in the opponents' rationality.

Consider a type $t_i^{c_i} \in T_i^*$ where $c_i \in C_i^*$. By (3.6.6), type $t_i^{c_i}$ only assigns positive probability to opponent's choice-type pairs $(c_j, t_j^{c_j})$ where $c_j \in C_j^*$. Since c_j is optimal for $t_j^{c_j}$, the type $t_i^{c_i}$ only assigns positive probability to opponent's choice-type pairs $(c_j, t_j^{c_j})$ where c_j is optimal for $t_j^{c_j}$. Hence, $t_i^{c_i} \in T_i^*$ believes in the opponents' rationality. Since this holds for every type $t_i^{c_i} \in T_i^*$, all types in M^* believe in the opponents' rationality. Hence, by Theorem 3.3.1, all types in M^* express common belief in rationality. Note that the types in M^* are exactly the types $t_i^{c_i}$ with $c_i \in C_i^K$. Hence, for every $c_i \in C_i^K$, the associated type $t_i^{c_i}$ expresses common belief in rationality.

We can now prove that $C_i^{k+1} \subseteq BR_i^k$ for all players i and all $k \geq 1$. Take some $c_i \in C_i^{k+1}$ where $k \geq 1$. Then we know from above that c_i is optimal for the associated type $t_i^{c_i}$, and that the type $t_i^{c_i}$ expresses up to k-fold belief in rationality. Hence, by definition, $c_i \in BR_i^k$. As this holds for every $c_i \in C_i^{k+1}$, we conclude that $C_i^{k+1} \subseteq BR_i^k$ for all players i and all $k \geq 1$.

Since in part (i) we have already seen that $BR_i^k \subseteq C_i^{k+1}$, we may conclude that $BR_i^k = C_i^{k+1}$ for all players i and all $k \geq 1$. That is, a choice can rationally be made while expressing up to k-fold belief in rationality precisely when the choice survives $k+1$ elimination rounds. This establishes part (a) of Theorem 3.4.1.

We finally prove part (b) of Theorem 3.4.1. Suppose first that choice c_i can rationally be made under common belief in rationality. Then, in particular, for every $k \geq 1$, the choice c_i can rationally be made while expressing up to k-fold belief in rationality. By part (a) we then know that c_i survives $k+1$ rounds of elimination. Since this holds for every $k \geq 1$, we conclude that c_i survives all rounds of elimination.

Suppose next that the choice c_i survives all rounds of elimination. Then, $c_i \in C_i^K$, where K is the round at which the procedure of *iterated elimination of strictly dominated choices* terminates. From the construction of the epistemic model $M = (T_i, b_i)_{i \in I}$ above we know that the choice c_i is optimal for the type $t_i^{c_i}$ and that the type $t_i^{c_i}$ expresses common belief in rationality. Hence, c_i can rationally be made under common belief in rationality. We thus conclude that a choice c_i can rationally be made under common belief in rationality precisely when the choice c_i survives all rounds of elimination. This completes the proof of part (b), and thereby the proof of this theorem. ■

Proof of Theorem 3.4.2. We know that for every player there is at least one choice that survives the procedure of *iterated elimination of strictly dominated choices*. For every player i, let C_i^* be the set of choices for player i that survive all rounds of the procedure, which contains at least one choice. Then, we can construct the epistemic model $M = (T_i, b_i)_{i \in I}$ as we did in the proof of Theorem 3.4.1 above. Consider the smaller epistemic model $M^* = (T_i^*, b_i)_{i \in I}$ contained in M in which

$$T_i^* := \{t_i^{c_i} \mid c_i \in C_i^*\},$$

and the beliefs of the types are given by (3.6.6). We already saw above that this is a well-defined epistemic model, and that all types in M^* express common belief in rationality. This completes the proof. ■

3.6.3 Proof of Section 3.5

Before we prove Theorem 3.5.1, we first introduce some additional definitions and results. We start by defining general *reduction operators,* which assign to an object a smaller object by eliminating parts

of it. As a next step, we define *elimination orders* for a given reduction operator, and state what it means that the final outcome is *independent* of the elimination order. We then provide a condition on reduction operators, called *monotonicity,* which guarantees that the final outcome is independent of the elimination order.

Next, it is shown how the iterated elimination of strictly dominated choices can be identified with the iterated application of a specific reduction operator. Finally, we show that the reduction operator which characterizes the iterated elimination of strictly dominated choices is monotone. As a consequence, the final outcome is independent of the elimination order, thus establishing Theorem 3.5.1.

3.6.3.1 Reduction Operators

Consider a finite set A. A *reduction operator* r on A assigns to every subset $D \subseteq A$ a smaller set $r(D) \subseteq D$. For a given set $D \subseteq A$ and a number $k \in \{1, 2, 3, ...\}$, we denote by $r^k(D) := \underbrace{r(r(...(r(D))...))}_{k \text{ times}}$ the k-fold application of the reduction operator r to D.

An *elimination order* for r is a finite sequence $(D^0, D^1, ..., D^K)$ where (a) $D^0 = A$, (b) $r(D^k) \subseteq D^{k+1} \subseteq D^k$ for every $k \in \{0, ..., K-1\}$, and (c) $r(D^K) = D^K$. Here, condition (a) states that we start with the full set A, condition (b) states that in round $k+1$ we eliminate at most as much from D^k as is allowed by r, but possible less, whereas condition (c) guarantees that D^K cannot be reduced any further, and hence the elimination procedure terminates there.

In an elimination order we thus start with the full set A, and during every consecutive round we eliminate from the current set at most as much as is allowed by r, but possible less, until we reach a point where the set cannot be reduced any further.

One special elimination order is the *full speed* elimination order $(D^0, D^1, ..., D^K)$, where $D^{k+1} = r(D^k)$ for every $k \in \{0, 1, ..., K-1\}$. In this elimination order, we always eliminate as much as possible in every round.

3.6.3.2 Order Independence

Some reduction operators have the special property that the final outcome will always be the same, no matter which elimination order is chosen. This property is called *order independence.*

Definition 3.6.1 (Order independence) *A reduction operator r is **order independent** if for every two elimination orders $(D^0, D^1, ..., D^K)$ and $(E^0, E^1, ..., E^L)$ we have that $D^K = E^L$.*

We will now introduce a condition, called *monotonicity,* which guarantees that the reduction operator is order independent.

Definition 3.6.2 (Monotonicity) *A reduction operator r is **monotone** if for every two sets D, E where $r(E) \subseteq D \subseteq E$, we have that $r(D) \subseteq r(E)$.*

Hence, monotonicity reveals the idea that smaller sets should have smaller reductions. The following result shows that monotonicity implies order independence.

Lemma 3.6.2 (Monotonicity implies order independence) *Every monotone reduction operator is order independent.*

Proof. Take a reduction operator r that is monotone, and a consider an arbitrary elimination order $(D^0, D^1, ..., D^M)$. For some $m \in \{0, 1, ..., M-1\}$, consider the sets D^m and D^{m+1}. We show the following property.

Claim. For every $k \geq 1$, we have that $r^{k+1}(D^m) \subseteq r^k(D^{m+1}) \subseteq r^k(D^m)$.

Proof of claim. We prove the statement by induction on k. We start with $k = 1$. As $(D^0, D^1, ..., D^M)$ is an elimination order for r, we know that $r(D^m) \subseteq D^{m+1} \subseteq D^m$. By monotonicity of r it then follows that $r(D^{m+1}) \subseteq r(D^m)$. We thus have that $r(D^{m+1}) \subseteq r(D^m) \subseteq D^{m+1}$. By applying monotonicity of r again, we conclude that $r(r(D^m)) \subseteq r(D^{m+1})$. Altogether, we see that $r^2(D^m) \subseteq r(D^{m+1}) \subseteq r(D^m)$, which proves the statement for $k = 1$.

Take now some $k \geq 2$, and assume that the statement is true for $k-1$. That is, we know that $r^k(D^m) \subseteq r^{k-1}(D^{m+1}) \subseteq r^{k-1}(D^m)$. Since $r^k(D^m) = r(r^{k-1}(D^m))$, it follows by monotonicity of r that $r(r^{k-1}(D^{m+1})) \subseteq r(r^{k-1}(D^m))$, and hence $r^k(D^{m+1}) \subseteq r^k(D^m)$.

Thus, we know that $r^k(D^{m+1}) \subseteq r^k(D^m) \subseteq r^{k-1}(D^{m+1})$, where $r^k(D^{m+1}) = r(r^{k-1}(D^{m+1}))$. By monotonicity of r, we then conclude that $r(r^k(D^m)) \subseteq r(r^{k-1}(D^{m+1}))$, and hence $r^{k+1}(D^m) \subseteq r^k(D^{m+1})$. Altogether, we see that $r^{k+1}(D^m) \subseteq r^k(D^{m+1}) \subseteq r^k(D^m)$, which establishes the statement for k.

By induction on k, the statement in the claim holds for every k. This completes the proof of the claim.

Consider now an arbitrary elimination order $(D^0, D^1, ..., D^M)$ for r. Let K be large enough such that $r^{K+1}(D^m) = r^K(D^m)$ for all $m \in \{0, 1, ..., M\}$. By applying the claim for $m = 0$, we get that $r^{k+1}(D^0) \subseteq r^k(D^1) \subseteq r^k(D^0)$ for all k. In particular, $r^{K+1}(D^0) \subseteq r^K(D^1) \subseteq r^K(D^0)$. As $r^{K+1}(D^0) = r^K(D^0)$, it must be that $r^K(D^1) = r^K(D^0)$. By applying the claim for all $m \in \{1, ..., M-1\}$, we conclude in a similar way that $r^K(D^{m+1}) = r^K(D^m)$ for all $m \in \{1, ..., M-1\}$. Hence, it follows in particular that $r^K(D^M) = r^K(D^0)$. As $r(D^M) = D^M$, we have that $r^K(D^M) = D^M$ as well, and hence $D^M = r^K(D^0) = r^K(A)$. Since this holds for every elimination order $(D^0, D^1, ..., D^M)$ we conclude that $D^M = r^K(A)$, independent of the specific elimination order chosen. Hence, r is order independent. This completes the proof. ■

3.6.3.3 Iterated Elimination of Strictly Dominated Choices

We now show how the iterated elimination of strictly dominated choices can be viewed as the iterated application of a specific reduction operator. For a given game, let $A = (C_i, C_{-i}, u_i)_{i \in I}$ be the set that assigns to every player i the (full) decision problem (C_i, C_{-i}, u_i). Since the utility function u_i is fixed, we just write $A = (C_i, C_{-i})_{i \in I}$ as an abbreviation. Consider two reduced decision problems (D_i, D_{-i}) and (E_i, E_{-i}), where $D_i, E_i \subseteq C_i$, and $D_{-i}, E_{-i} \subseteq C_{-i}$. We say that $(D_i, D_{-i}) \subseteq (E_i, E_{-i})$ if $D_i \subseteq E_i$ and $D_{-i} \subseteq E_{-i}$. Similarly, for two subsets $D = (D_i, D_{-i})_{i \in I}$ and $E = (E_i, E_{-i})_{i \in I}$ of A, we write $D \subseteq E$ if $(D_i, D_{-i}) \subseteq (E_i, E_{-i})$ for every player i.

Let sd be the reduction operator that assigns to every set of decision problems $E = (E_i, E_{-i})_{i \in I}$ a smaller set of decision problems $sd(E) = (D_i, D_{-i})_{i \in I}$ where, for all players i,

$$\begin{aligned} D_{-i} &:= \{(c_j)_{j \neq i} \in E_{-i} \mid c_j \in E_j \text{ for all } j \neq i\} \text{ and} \\ D_i &:= \{c_i \in E_i \mid c_i \text{ not strictly dominated in } (E_i, D_{-i}, u_i)\}. \end{aligned} \tag{3.6.7}$$

This reduction operator is called the *strict dominance reduction operator.*

For every round k, let C_i^k and C_{-i}^k be the set of choices and the set of states that survive round k of the iterated elimination of strictly dominated choices in i's decision problem. In the light of (3.6.7),

we then conclude that

$$sd^k(A) = (C_i^k, C_{-i}^k)_{i \in I} \tag{3.6.8}$$

for every round k. In other words, the iterated elimination of strictly dominated choices is equivalent to the iterated application of the strict dominance reduction operator at full speed.

Lemma 3.6.3 (Strict dominance operator is monotone) *The strict dominance reduction operator is monotone.*

Proof. Consider two sets of decision problems $D = (D_i, D_{-i})_{i \in I}$ and $E = (E_i, E_{-i})_{i \in I}$ where $sd(E) \subseteq D \subseteq E$. We must show that $sd(D) \subseteq sd(E)$. Let $sd(D) = (D_i', D_{-i}')_{i \in I}$ and $sd(E) = (E_i', E_{-i}')$. Hence, we must show, for every player i, that $D_i' \subseteq E_i'$ and $D_{-i}' \subseteq E_{-i}'$.

We start by showing that $D_{-i}' \subseteq E_{-i}'$. Take some $(c_j)_{j \neq i}$ in D_{-i}'. Then, by definition of the sd operator, $c_j \in D_j$ for every $j \neq i$. As $D \subseteq E$ we have that $D_j \subseteq E_j$ for all $j \neq i$, and hence $c_j \in E_j$ for all $j \neq i$. Thus, by definition of the sd operator, $(c_j)_{j \neq i}$ is in E_{-i}'. This shows that $D_{-i}' \subseteq E_{-i}'$.

We next prove that $D_i' \subseteq E_i'$. Take some $c_i \in D_i'$. By definition of the sd operator, we have that c_i is not strictly dominated in (D_i, D_{-i}', u_i). By Theorem 2.7.1 it follows that there is some belief $b_i \in \Delta(D_{-i}')$ such that

$$u_i(c_i, b_i) \geq u_i(c_i', b_i) \text{ for all } c_i' \in D_i. \tag{3.6.9}$$

Note that $b_i \in \Delta(E_{-i}')$ since we have seen that $D_{-i}' \subseteq E_{-i}'$. Now, let $c_i^* \in E_i$ be such that

$$u_i(c_i^*, b_i) \geq u_i(c_i', b_i) \text{ for all } c_i' \in E_i. \tag{3.6.10}$$

By Theorem 2.7.1, we conclude that c_i^* is not strictly dominated in (E_i, E_{-i}', u_i), and hence $c_i^* \in E_i'$ by definition of the sd operator. Since $sd(E) \subseteq D$, we know, in particular, that $E_i' \subseteq D_i$, and thus we see that $c_i^* \in D_i$. By combining (3.6.9) and (3.6.10), and using the fact that $c_i^* \in D_i$, we conclude that

$$u_i(c_i, b_i) \geq u_i(c_i^*, b_i) \geq u_i(c_i', b_i) \text{ for all } c_i' \in E_i.$$

By Theorem 2.7.1 it then follows that c_i is not strictly dominated in (E_i, E_{-i}', u_i), and hence c_i is in E_i'. This shows that $D_i' \subseteq E_i'$.

Altogether, we see that $D_{-i}' \subseteq E_{-i}'$ and $D_i' \subseteq E_i'$, and thus $sd(D) \subseteq sd(E)$. Since this holds for every two sets of decision problems $D = (D_i, D_{-i})_{i \in I}$ and $E = (E_i, E_{-i})_{i \in I}$ where $sd(E) \subseteq D \subseteq E$, we conclude that the strict dominance reduction operator sd is monotone. This completes the proof. ■

We are now ready to prove Theorem 3.5.1.

Proof of Theorem 3.5.1. We have seen above that the iterated elimination of strictly dominated choices is equivalent to the iterated application of the strict dominance reduction operator sd at full speed. Moreover, we know by Lemma 3.6.3 that the sd operator is monotone. Together with Lemma 3.6.2, we conclude that sd is order independent. Hence, the final outcome of the iterated elimination of strictly domnated choices does not depend on the specific elimination order we use. This completes the proof. ■

Solutions to In-Chapter Questions

Question 3.1.1. At the belief $p = (0.6)\cdot$ *blue* $+ (0.4)\cdot$ *green,* the expected utilities of the four colors are

$$\begin{aligned} u(\textit{blue},\, p) &= (0.6)\cdot 0 + (0.4)\cdot 4 = 1.6, \\ u(\textit{green},\, p) &= (0.6)\cdot 3 + (0.4)\cdot 0 = 1.8, \\ u(\textit{red},\, p) &= (0.6)\cdot 2 + (0.4)\cdot 2 = 2, \text{ and} \\ u(\textit{yellow},\, p) &= (0.6)\cdot 1 + (0.4)\cdot 1 = 1. \end{aligned}$$

Hence, your preference relation at that belief is *red* $\succ_p$ *green* $\succ_p$ *blue* $\succ_p$ *yellow.*

Question 3.1.2. It may be verified that the randomized choice $(0.5)\cdot$ *blue* $+ (0.5)\cdot$ *green* strictly dominates your choice *yellow* under the expected utility representation u from Table 3.1.1.

To see that it strictly dominates *yellow* for *every* expected utility representation, note first that there are preference reversals between every pair of choices. Moreover, there are beliefs where you are indifferent between some, but not all, colors. We thus know by Theorem 2.5.2 that the relative preference intensities are unique. But then, the randomized choice $(0.5)\cdot$ *blue* $+ (0.5)\cdot$ *green* strictly dominates *yellow* for *every* expected utility representation.

Question 3.2.1. If we start at your choice *green,* then you believe that Barbara assigns probability 0.6 to you choosing *red* and that Barbara assigns probability 0.4 to you choosing *yellow.* If we start at Barbara's choice *green,* then Barbara believes that you believe that Barbara chooses *blue.* If we start at Barbara's choice *blue,* then Barbara's assigns probability 0.6 to the event that you believe that Barbara chooses *blue* and *green* with probabilities 0.6 and 0.4, respectively, and Barbara assigns probability 0.4 to the event that you believe that Barbara chooses *yellow.*

Question 3.2.2. If we start at your choice *blue,* then you believe that Barbara believes that you believe that Barbara chooses *red.* If we start at Barbara's choice *green,* then Barbara believes that you believe that Barbara assigns probability 0.6 to you choosing *red* and probability 0.4 to you choosing *yellow.*

Question 3.2.3. Suppose we delete the arrow from your choice *blue* to Barbara's choice *red*, and the arrow from Barbara's choice *red* to your choice *blue.* Then, the only choices with outgoing arrows would be your choices *green, red* and *yellow,* and Barbara's choices *blue, green* and *yellow.* Moreover, every choice with an outgoing arrow leads only to choices with outgoing arrows, and hence this qualifies as a beliefs diagram. If, in addition, we delete the arrow from Barbara's choice *green* to your choice *green,* then the forked arrow starting at your choice *red* would lead, with probability 0.4, to Barbara's choice *green,* which does not have an outgoing arrow. Hence, we cannot derive the full belief hierarchy if we start at your choice *red,* and this would therefore not qualify as a beliefs diagram.

Question 3.2.4. Under this probabilistic belief, Barbara's expected utility from choosing *blue, green, red* and *yellow* are

$$u_2(\textit{blue}) = 0,\ u_2(\textit{green}) = (0.3)\cdot 4 = 1.2,\ u_2(\textit{red}) = (0.7)\cdot 1 = 0.7 \text{ and } u_2(\textit{yellow}) = 2.$$

Hence, *yellow* yields Barbara the highest expected utility under this belief.

Question 3.2.5. The first-order belief about Barbara is that you think that Barbara chooses *green.* In your second-order belief, you believe that Chris believes that Barbara chooses *yellow.* In your

third-order belief, you believe that Chris believes that Barbara assigns probability 0.3 to the event that you choose *red* and Chris chooses *blue,* and that Barbara assigns probability 0.7 to the event that you choose *green* and Chris chooses *blue.*

Question 3.2.6. Your type t_1^{blue} believes that Barbara chooses *red* (first-order belief), that Barbara believes that you choose *blue* (second-order belief), and that Barbara believes that you believe that Barbara chooses *red* (third-order belief). Barbara's type t_2^{green} believes that you choose *green* (first-order belief), that you believe that Barbara chooses *blue* (second-order belief), and that you believe that Barbara assigns probability 0.6 to you choosing *red* and probability 0.4 to you choosing *yellow* (third-order belief).

Question 3.2.7. You believe that Barbara chooses *blue* and Chris chooses *yellow.* You believe that Barbara believes that Chris chooses *yellow.* You believe that Chris believes that you believe that Barbara chooses *blue.*

Question 3.3.1. For you, *blue* is optimal for t_1^{blue}, *green* is optimal for t_1^{green}, *red* is optimal for t_1^{red} and *blue* is optimal for t_1^{yellow}. For Barbara, *blue* is optimal for t_2^{blue}, *red* is optimal for t_2^{green}, *red* is optimal for t_2^{red} and *yellow* is optimal for t_2^{yellow}.

Question 3.3.2. Your types t_1^{blue}, t_1^{green} and t_1^{yellow} believe in Barbara's rationality. Barbara's types t_2^{green}, t_2^{red} and t_2^{yellow} believe in your rationality. Barbara's type t_2^{blue} does not believe in your rationality since it assigns positive probability to your choice-type pair $(yellow, t_1^{yellow})$, whereas *yellow* is notoptimal for your type t_1^{yellow}.

Question 3.3.3. Barbara's type t_2^{yellow} believes that you hold the type t_1^{red}, which does not express 1-fold belief in rationality.

Question 3.3.4. Barbara's type t_2^{green} believes that you hold the type t_1^{green}, which does not express 2-fold belief in rationality.

Question 3.5.1. Here is an example: In round 1, only eliminate your choice *yellow.* In round 2, only eliminate Barbara's choice *green.* In round 3, only eliminate your choice *red.* In round 4, only eliminate Barbara's choice *blue.* In round 5, only eliminate your choice *green.* In round 6, only eliminate Barbara's choice *yellow.* Then, only your choice *blue* and Barbara's choice *red* survive. It takes 6 rounds to terminate. There are also other orders of elimination where only one choice is eliminated in every round. However, each of these elimination orders takes 6 rounds to terminate, and the surviving choices will always be the same as here.

Problems

Problem 3.1: The dancing competition.

You and Barbara are the jury members of a dancing competition. After every performance, you and Barbara report an integer grade between 4 and 10, and the final grade for the candidate will just be the average of the two reported grades.

You have just seen the first dancer, and you value her performance at 8, whereas Barbara values the performance at 7. That is, your desired grade is an 8 and Barbara's desired grade is a 7. Since you are both loud and clear about your opinion, you both know the desired grade of the other person. Now it is time for both of you to report a grade, which may well be different from your desired grade, because the final grade also depends on the other person's reported grade.

Suppose that the conditional preference relation for you is such that the closer the final grade is to your desired grade, the higher your utility. More precisely, for every pair of reported grades your utility is 4 minus the *absolute* difference between your desired grade and the final grade. For instance, if you report a grade of 9 and Barbara a grade of 5, then the final grade is 7, and thus your utility is $4 - |8 - 7| = 4 - 1 = 3$. If you both report a grade of 10, then the final grade would be 10, and your utility would be $4 - |8 - 10| = 4 - 2 = 2$. The conditional preference relation for Barbara is similar, but recall that her desired grade is a 7 instead of an 8.

(a) Formulate this story as a game, by specifying the decisions problems for you and Barbara.

(b) Which choices (reported grades) are rational for you? For every rational choice, find a belief about Barbara's choice for which that choice is optimal. For every irrational choice, find another choice, or randomized choice, that strictly dominates it.

(c) Which grades can you and Barbara rationally report while expressing up to 3-fold belief in rationality? Which grades can you and Barbara rationally report under common belief in rationality? Which final grades are possible under common belief in rationality?

Barbara was shocked by the large difference in grades you both reported for the first dancer, and it has made her rather insecure. Her preferences have changed because of this unpleasant experience. From now on, her objective is to report the same grade as you. That is, if she reports the same grade as you, her utility will be 1, whereas her utility will be 0 if she reports a grade different from yours. Your conditional preference relation is similar as before: Your utility is 4 minus the absolute difference between your desired grade and the final grade.

The second dancer in the competition is Chris, who danced the jive just like John Travolta in Saturday Night Fever. Your desired grade for Chris' marvellous performance is therefore a 9.

(d) Which grades can you and Barbara rationally report under common belief in rationality? Which final grades are possible under common belief in rationality?

(e) Make a beliefs diagram with solid arrows only that includes precisely the choices you found in (d), but no other choices. Which of your belief hierarchies in the diagram expresses common belief in rationality?

(f) Translate this beliefs diagram into an epistemic model. Which of the types express common belief in rationality? Make sure that for every choice found in (d) there is a type that expresses common belief in rationality and for which that choice is optimal.

Problem 3.2: Selling ice cream on the beach.

Barbara, Chris and you are spending a lovely holiday on an island with many beaches. Today, you all want to sell ice cream on the beach. There are five beaches close to the hotel: Acapulco Beach with room for 1400 people, Bounty Beach with room for 1200 people, Cucaracha Beach with room for 1000 people, Dream Beach with room for 900 people, and El Dorado Beach with room for 600 people. You expect all beaches to be packed today, since the weather is absolutely fantastic. Moreover, everybody at the beach is expected to buy exactly one ice cream.

In the early morning, Barbara, Chris and you must independently decide to which beach you will go to sell your ice cream. If you happen to go to the same beach as one of your friends, then you will both sell to half of the people at that beach. If you all go to the same beach, then each of you will sell to one third of the people at that beach. Of course, the objective for each of you is to sell as many ice creams as possible. As such, the utility you obtain will always be equal to the number of customers to which you sell.

(a) How many states does your decision problem contain?

(b) In (a) you have seen that there are many states – too many to write down your full decision problem in a reasonable amount of time. To at least get an idea of how your decision problem looks like, write down the *reduced* decision problem for you in which you can choose each of the five possible locations, but where the states only contain Acapulco Beach, Bounty Beach and Cucaracha Beach for Barbara and Chris.

(c) Which beaches are rational in your decision problem, and which are irrational? For every rational beach, find a belief about Barbara's and Chris' choice for which that beach is optimal. For every irrational beach, find another beach, or a randomized choice, that strictly dominates it.

(d) To which beaches can you rationally go under common belief in rationality?

(e) Construct a beliefs diagram with solid arrows only that uses, for every player, precisely those choices found in (d).

(f) Consider the belief hierarchy that supports your choice *Cucaracha Beach.* In your first-order belief, what do you believe that Chris will do? In your second-order belief, what do you believe that Chris believes that Barbara will do? Answer the same two questions for the belief hierarchy that supports your choice *Dream Beach.*

(g) Translate this beliefs diagram into an epistemic model, where all types express common belief in rationality, and where for every choice found in (d) there is a type for which that choice is optimal.

During breakfast Chris told you, quite surprisingly, that he will work as a life guard at Cucaracha Beach today, and will therefore not be able to sell ice creams.

(h) To which beaches can you rationally go under common belief in rationality within this new scenario? Compare your findings with (d), and explain the difference.

Problem 3.3: Deborah's garden.

You have a friend, Deborah, with a magnificent garden. Tomorrow, she needs someone to help her getting the garden ready for the evening party. Barbara, Chris and you are all interested in helping her, but she only needs one person she said. To decide who will help her, you must all write a price on a piece of paper and give it to Deborah. The person who writes down the lowest price will help her tomorrow, and gets paid exactly the amount he or she wrote down. You can only write down prices of 10 euros, 20 euros, 30 euros, up to 100 euros. If two persons write down the same lowest price, then

Deborah will toss a coin to decide who gets to help her. If you all three write down the same price, she will throw a dice to determine the garden assistant. Of course, in that case you all have an equal chance to become her assistant.

Assume that your conditional preference relation is represented by a utility function where, for every price of yours, and every combination of prices by Barbara and Chris, your utility equals the expected reward you obtain. Similarly for Barbara and Chris.

(a) How many states are there in your decision problem?

(b) You have seen in (a) that there are many states, and therefore writing down your complete decision problem is a very cumbersome and boring task. To get an idea of how your decision problem would look like, write down the *reduced* decision problem that only contains your choices $100, 90$ and 10, and that only contains states where Barbara and Chris choose $100, 90$ or 80.

(c) Which prices are rational and which are irrational in your decision problem? For every rational price, find a belief about Barbara's and Chris' choice for which that price is optimal. For every irrational price, find another price, or randomized choice, that strictly dominates it.

(d) What prices can you rationally choose while expressing up to 4-fold belief in rationality? What prices can you rationally choose under common belief in rationality?

It is now one day later, and Deborah again needs help from somebody, this time to help her clean up the garden after the party. She proposes the same procedure as above. However, since Barbara, Chris and you were very disappointed by the outcome of the procedure yesterday, you have decided to do things differently this time. You have agreed that the person who is selected to help Deborah will share his or her income equally with the other two friends. For instance, if Barbara is selected to help Deborah at a price of 60 euros, everybody will obtain 20 euros at the end.

(e) What prices can you rationally choose under common belief in rationality?

(f) Construct an epistemic model where all types express common belief in rationality, and such that for every price found in (e) there is a type for which that price is optimal.

A choice in a game is called *weakly dominant* if, whatever the opponents do, this choice is always at least as good than any other choice, and sometimes strictly better.

(g) Find the unique weakly dominant price for you, Barbara and Chris in this new scenario. Compare this to the outcome in (d), and explain the difference.

Literature

Origins of game theory. Why is the field called "game theory" and not "multi-person decision making under uncertainty", for instance, which would be a more accurate description of its content? The reason, as often, is historical. Some of the early developments, which later gave rise to the field, were explicitly about recreational games.

More than a century ago, Zermelo (1913) proved that in the game of *chess*, every position is either a winning position for White, or a winning position for Black, or a non-loosing position for both White and Black. As a consequence, either White has a strategy that guarantees him a win, no matter what Black does, or Black has a strategy that guarantees him a win, no matter what White does, or both White and Black have a strategy that guarantee them at least a draw, no matter what the opponent does. This result, which often is referred to as Zermelo's Theorem, is generally regarded as the first theorem in game theory,

During the next decade, Borel (1921, 1924, 1927) investigated *symmetric two-person zero-sum games* involving chance, which are recreational games with two players where both players have symmetric roles, and where the gain of x for one player results in the loss of the same amount x for the other player. Borel starts his analysis with the iterated elimination of *bad* strategies, which are strategies that give the player an expected payoff of at most 0. In Borel (1921) it is shown that, if the reduced game obtained at the end contains three strategies for both players, then each player has a randomization over strategies that gives him an expected payoff of exactly 0, no matter what the opponent plays. The follow-up papers Borel (1924) and Borel (1927) extend this result to the case where the reduced game contains five and seven strategies, respectively.

Borel's findings were later generalized by von Neumann (1928) who studied general two-person zero-sum games in which the players may have asymmetric roles. He proved that for every such game there is a unique number v such that (a) player 1 has a randomization over strategies that gives him an expected payoff of at least v, no matter what player 2 does, and (b) player 2 has a randomization over strategies that gives him an expected payoff of at least $-v$, no matter what player 1 does. This number v is called the *value* of the zero-sum game, and this result is known as von Neumann's *maxmin theorem.* Clearly, if the game is symmetric, as Borel assumes, than the value must be 0, and hence Borel's result follows.

Later it was recognized that the essential ingredients of a recreational game – that there are various players involved, that every player chooses a strategy for how to play the game, and that the final outcome depends on the strategies of all players – are also present in many scenarios in economics, politics, and other environments where human decision making plays an important role. In that light, the book *Theory of Games and Economic Behavior* by von Neumann and Morgenstern (1944) was a true milestone, as it showed how many of such scenarios in economic theory can be modelled, and analyzed, in a uniform way. The book thereby gave birth to the field of game theory as a scientific discipline. The reader who wants to learn more about the influence that John von Neumann and Oskar Morgenstern had on the creation of game theory may consult the book by Leonard (2010).

Games. The first to give a general and systematic definition of a game was von Neumann (1928). He defined a dynamic game in which the players may have to make a sequence of choices during a finite number of periods. At every period the players simultaneously make a choice, which moves the game to the next period. He also allows for chance moves with commonly known objective probabilities. If we take von Neumann's definition and apply it to a scenario with one period only, we essentially obtain the definition of a game in this chapter. The only difference is that we do not involve chance

moves in our definition. In their book, von Neumann and Morgenstern (1944) build on von Neumann's (1928) definition.

Games as decision problems. We have seen that a game can be viewed as a combination of *decision problems* – one for every player – in which the states correspond to the possible opponents' choice combinations. As such, we explicitly take a *one-person perspective* in this book. This is also reflected by the definitions, examples and exercises in which we always take the viewpoint of a single player – be it player i or "you" – and reason about the game from the perspective of this single player. Such a one-person approach to game theory is also present in Harsanyi (1967–1968) when he lays out his framework for games with incomplete information, in which players are uncertain about the opponents' utility functions.

Belief hierarchies. In a sense, belief hierarchies constitute the language of epistemic game theory. They may be viewed as the end-product of a reasoning process that results in a belief about the opponents' choices, a belief about the opponents' beliefs about the other players' choices, and so on.

The importance of belief hierarchies for economic theory has already been stressed by Morgenstern (1935). In this paper he criticizes the common assumption in economics that all agents are correct about everything in the model – the parameters of the model, but also the behavior of other agents. But if this correctness assumption is dropped, it becomes important to model what an agent *believes* about the parameters and the behavior of other agents, what an agent believes about the beliefs of other agents, and so on.

Despite this, belief hierarchies have remained absent from game theory for many decades after Morgenstern's article. A possible reason is that the early achievements by Zermelo (1913), Borel (1921, 1924, 1927), von Neumann (1928) and Nash (1950, 1951), but also the book by von Neumann and Morgenstern (1944), have pushed game theory in a direction where the need for belief hierarchies was heavily diminished. To see why, consider the early results by Zermelo, Borel and von Neumann as discussed above. In each of these theorems, the focus is on a player who can choose a certain (randomized) strategy that guarantees him a particular (expected) outcome, *no matter what the opponent does*. Hence, to guarantee this outcome the player need not actively reason about the opponent's choices or beliefs, as the strategy is guaranteed to deliver this outcome, or more, independent of what the opponent does. But if there is no need to reason about the opponent, the role of belief hierarchies, as mathematical representations of the players' reasoning, is heavily diminished. Later, the book by von Neumann and Morgenstern (1944) adopted this "reasoning-free" approach to games, despite Morgenstern's (1935) arguments in favor of reasoning and belief hierarchies. As we will see in the following chapter, Nash's (1950, 1951) influential equilibrium concept requires every player i to believe that his opponents are *correct* about i's beliefs, thereby imposing a *correctness assumption* similar to the one criticized by Morgenstern (1935). This, again, led to an analysis that avoids the explicit use of reasoning and belief hierarchies. Because Nash equilibrium would play a dominant role in game theory for many decades, reasoning and belief hierarchies remained absent from the game-theoretic picture for a long time.

To the best of my knowledge, Harsanyi (1962) was the first to explicitly incorporate belief hierarchies into game theory, although he did so for the very specific context of bargaining situations between two persons who face uncertainty about the opponent's utility function. The belief hierarchies he explored described the belief of party 1 about the best terms that party 2 is willing to accept, the belief of party 1 about party 2's belief about the best terms that party 1 is willing to accept, and so on. He also restricted attention to probability 1 beliefs. Later, Harsanyi (1967–1968) extended the use of belief hierarchies to general *games with incomplete information* (see Chapters 5 and 6 in this book) in which players have uncertainty about the opponents' utility functions. For such games, he allowed

for *probabilistic* belief hierarchies describing what a player believes about the opponents' choices and utility functions, what he believes about the opponents' beliefs about the other players' choices and utility functions, and so on.

Since Harsanyi showed how belief hierarchies can be incorporated into the analysis of games, he deserves much of the credit for the transition from classical to epistemic game theory. The reader who wants to know more about this transition, and why it took so long, may consult the historical overview papers by Brandenburger (2010) and Perea (2014).

Beliefs diagrams. Beliefs diagrams have been introduced in the textbook Perea (2012) as a visual representation of belief hierarchies.

Types. The use of types to mathematically encode belief hierarchies goes back to Harsanyi (1967–1968). For games with incomplete information, where players face uncertainty about the opponents' utility functions, Harsanyi introduced epistemic models with types to represent belief hierarchies in a compact and convenient way. In Harsanyi's framework, every type for a player prescribes a utility function, a choice, and a probabilistic belief about the opponents' types. From this model we can then derive, for every type, a full infinite belief hierarchy about the choices and utility functions for the players, similarly to how we have derived belief hierarchies from types in this chapter.

The main difference with Harsanyi's model is that we do not prescribe a choice for a type. In turn, the types in our model hold a belief about the opponents' types *and the opponents' choices,* instead of having a belief solely about the opponents' types. This is necessary in order to be able to derive a belief hierarchy on choices from a type. Another difference, of course, is that we do not prescribe a utility function for a type, since in this chapter we concentrate on games with *complete* information in which players are informed about the opponents' utility functions.

In Harsanyi (1967–1968), special attention is paid to a scenario where the probabilistic beliefs about the opponents' types are derived from a *common prior* probability distribution on the type combinations by the players. This assumption is often called the *Harsanyi doctrine.* We will come back to the common prior in the next chapter.

The epistemic model we use in this chapter is essentially the one employed in Tan and Werlang (1988).

Alternative encodings of belief hierarchies. Encoding belief hierarchies by means of types *à la* Harsanyi is just one possible way of doing so. In the game-theoretic literature there are at least two important alternative ways for describing a belief hierarchy. The first approach, which is based on the models by Kripke (1963) and Aumann (1974, 1976), assumes that there is a set of *states of the world,* and a function that assigns to every state of the world a choice for each of the players. A player, however, has uncertainty about the true state of the world. This is modelled by assuming that at every state, there is a set of states – typically containing more than one state – which the player deems possible there. Moreover, at every state a player may hold a probabilistic belief about the states of the world he deems possible. In a similar way as for types, we can then derive a full infinite belief hierarchy about choices for every state and every player. We call this the *state-based approach.* The second approach, which is often used by scientists from logic and computer science, explicitly describes all levels in the belief hierarchy as formulae in some formal syntax. We call this the *syntactic* approach.

Among others, Brandenburger and Dekel (1993), Tan and Werlang (1992) and Bach and Perea (2021) explicitly compare the type-based approach, as we use it in this book, with the state-based approach described above. They show how to transform the encoding of a belief hierarchy in one model into an "epistemically equivalent" encoding in the other model, thereby establishing that the

two approaches are essentially equivalent. For the presentation and analysis of epistemic concepts it does not really matter which language one uses for describing the belief hierarchies – the type-based language, the state-based language or the syntactic language. What matters is the content of the belief hierarchies, and the conditions we impose on these. Hence, we could as well have written this book entirely by using one of the other two languages described above. In a sense, all these models just provide different representations of the same primitive notion, which is the belief hierarchy of a player.

Large epistemic models. In this chapter we have used epistemic models with types to encode certain belief hierarchies we are interested in. In a sense, for every belief hierarchy of interest we could construct a new epistemic model that encodes this particular belief hierarchy.

We say that the epistemic model is *terminal* if every possible belief hierarchy we can think of is already contained in this single epistemic model. Since there are obviously infinitely many – and in fact uncountably many – possible belief hierarchies, any terminal epistemic model must necessarily contain infinitely many – in fact, uncountably many – types for every player.

An important – but difficult – question that has been addressed in the literature is whether we can always construct a terminal epistemic model for every game. Armbruster and Böge (1979), Böge and Eisele (1979) and Mertens and Zamir (1985) were the first to explicitly construct such terminal epistemic models. Later, Brandenburger and Dekel (1993), Heifetz (1993) and Heifetz and Samet (1998) extended the above constructions by relaxing the topological assumptions being made in the model. Epstein and Wang (1996) show that a similar construction also works in a more general framework in which the players hold hierarchies of *preferences over acts*, rather than hierarchies of beliefs, satisfying certain regularity conditions.

The epistemic models we use in this book all contain finitely many types for every player, and are therefore necessarily not terminal. The reason we do not use terminal epistemic models is that we do not really need them for our purposes here. Moreover, epistemic models with finitely many types have the advantage that they can more easily be represented in examples – something that we find very important in this book.

Common knowledge and common belief. In the literature there is a distinction between *knowledge* and *belief*. The fundamental difference is that you can only know an event if that event is true, whereas you can believe an event which is not true. This is called the *truth axiom* of knowledge.

The notions of common knowledge and common belief have independently been defined by the sociologist Friedell (1967, 1969), the philosopher Lewis (1969) and the game theorist Aumann (1976). Both Friedell and Lewis use a syntactic approach, whereas Aumann employs a state-based approach for the definition.

Friedell uses the term *common opinion* rather than *common belief*, and he defines his notion in essentially the same way as how we define common belief in rationality in this chapter. He then defines *common knowledge* in an event as the situation where the event is a matter of common opinion between the persons involved, and where the event is true. Lewis (1969) defines common knowledge in a fundamentally different way, as he proceeds by identifying *sufficient conditions* that imply common opinion (knowledge) in the sense of Friedell. Aumann's (1976) definition of common knowledge is similar to Friedell's, but uses a state-based formulation instead of a syntactic one.

Common belief in rationality. As we have argued in this chapter, common belief in rationality is the central concept in epistemic game theory. The idea of common belief in rationality already appears in Friedell (1969) and Spohn (1982), although these two papers do not offer a fully rigorous definition. The definition of common belief in rationality as we use it in this chapter is taken from Tan

and Werlang (1988), who call it common *knowledge* of rationality instead. We have decided to use the term common belief, because we think it better fits the reasoning of people in game-like situations. A player in a game can never be fully certain of the opponents' choices or beliefs, and therefore knowledge seems too strong a concept for such scenarios.

In the literature, some other papers identify *sufficient conditions* that imply common belief in rationality. Examples are Harsanyi (1967–1968), Böge and Eisele (1979), Armbruster and Böge (1979), Aumann (1974, 1987) and Brandenburger and Dekel (1987). The first three papers consider Harsanyi-style models with types, and require that for every type the prescribed choice is optimal for the prescribed utility function and the induced belief about the opponents' choices. This property can be shown to imply common belief in rationality. The latter three papers above impose a similar condition in a state-based model.

The sufficient condition for common belief in rationality above is often called *universal rationality.* It is similar to the sufficient condition studied in this chapter, stating that every type in the epistemic model believes in the opponents' rationality. Indeed, both this sufficient condition and universal rationality state that there is *no irrationality in the system.* We have shown in Theorem 3.3.1 that under our sufficient condition, all types in the epistemic model express *common* belief in rationality. In a similar fashion it can be shown that also the universal rationality condition above implies common belief in rationality.

The idea of common belief in rationality is also *implicitly* present in the concept of *rationalizability,* as defined independently by Bernheim (1984) and Pearce (1984). Although these papers do not provide a formal definition of common belief in rationality, they argue informally that rationalizability is really based on this very idea.

Recursive procedure. In this chapter we have presented the recursive procedure known as the *iterated elimination of strictly dominated choices.* It is very similar to procedures that appear in Böge and Eisele (1979, Theorem 2), Armbruster and Böge (1979, Example 6.2), Pearce (1984, Definition 1) and Tan and Werlang (1988, Definition 5.1).

Our Theorem 2.7.1, which states that, within a decision problem, a choice is optimal for some belief precisely when it is not strictly dominated, is based on Pearce (1984, Lemma 3). This result is crucial for showing that the *iterated elimination of strictly dominated choices* characterizes the choices that can rationally be made under common belief in rationality.

Theorem 3.4.1, which shows that the *iterated elimination of strictly dominated choices* characterizes precisely those choices that can rationally be made under common belief in rationality, is perhaps *the* most important result in epistemic game theory. Brandenburger (2014) calls it the *fundamental theorem of epistemic game theory.* Different versions of this theorem can be found in Böge and Eisele (1979, Theorem 2), Brandenburger and Dekel (1987, Proposition 2.1) and Tan and Werlang (1988, Theorems 5.2 and 5.3). Spohn (1982) provides an intuition for why this theorem holds.

Order of elimination. In Section 3.5 we have seen that the order of elimination does not matter for the eventual output of the *iterated elimination of strictly dominated choices.* Papers that study the order independence of general, or specific, iterated elimination procedures are, for instance, Gilboa, Kalai and Zemel (1990), Apt (2004, 2011), Chen and Micali (2013), Luo, Qian and Qu (2020) and Perea (2017, 2018). The monotonicity condition we use for proving the order independence is very similar to the condition of *1-monotonicity** as proposed in Luo, Qian and Qu (2020).

Independent beliefs. In a game with three players or more, player i is said to have independent beliefs about the opponents' choices if for every two opponents j and k, his belief about opponent j's choice is stochastically independent from his belief about opponent k's choice. The concept of

rationalizability by Bernheim (1984) and Pearce (1984) assumes independent beliefs by the players. As this concept is also based on the idea of common belief in rationality, rationalizability is more restrictive than common belief in rationality. One way to see this more formally is that the recursive elimination procedure by Pearce (1984, Definition 1), which constitutes Pearce's definition of rationalizability, is more restrictive than the *iterated elimination of strictly dominated choices* which, we have seen, characterizes precisely those choices that can rationally be made under common belief in rationality.

Bernheim (1984) defends the independent beliefs assumption by arguing that the players typically make their choices *independently* from each other, without any possibility of communication, and that therefore a player's belief about the opponents' choices must be independent. In our view, this conclusion is not entirely correct: Even if player i believes that his opponents j and k choose independently, then it may still be that his belief about j's belief hierarchy is *correlated* with his belief about k's belief hierarchy. As a consequence, his belief about j's choice may well be correlated with his belief about k's choice, as these choices may arise as the optimal choices under different belief hierarchies, about which player i holds correlated beliefs.

This is precisely the viewpoint taken by Brandenburger and Friedenberg (2008), who call the above type of correlation between beliefs about different opponents' belief hierarchies *intrinsic.* They weaken the independence assumption in Bernheim (1984) and Pearce (1984) by stating that in a game with three players or more, the belief that player i has about opponent j's choice must be stochastically independent from his belief about opponent k's choice, once we *condition an a fixed belief hierarchy* for opponents j and k. They call this condition *conditional independence.* In other words, if we fix two belief hierarchies of opponents j and k, then conditional on these belief hierarchies being the actual belief hierarchies held by the opponents, the beliefs about the opponents' choices must be independent. This still reflects Bernheim's viewpoint that the two opponents choose independently, but recognizes that it does not automatically mean that the beliefs about the opponents' choices must be independent – only if we condition on some fixed belief hierarchies for the opponents it will. This condition is obviously weaker than the independence condition in Bernheim (1984) and Pearce (1984). Brandenburger and Friedenberg (2008) then take the concept of common belief in rationality and additionally impose common belief in the event that types have conditionally independent beliefs. The concept obtained is, in terms of choices selected, in between the concept of common belief in rationality and the concept of rationalizability.