

CHAPTER 9

Strong Belief in the Opponents' Rationality

9.1. Strong Belief in the Opponents' Rationality

In the previous chapter we have introduced the idea of *common belief in future rationality* for dynamic games, which means that you always believe that your opponents choose rationally now and in the future, you always believe that your opponents always believe that their opponents choose rationally now and in the future, and so on. This concept thus puts no restrictions on what you think about opponents' choices that have been made in the past – in fact you are free to conclude anything you want from what your opponents have done so far, as long as you still believe that these opponents will choose rationally now and in the future, that you still believe that these opponents also reason in this way about *their* opponents, and so on. So, the idea of common belief in future rationality makes a very sharp distinction between reasoning about the past and reasoning about the future: Anything goes for reasoning about the past, but very severe conditions are being imposed on how you reason about the future.

In many dynamic situations, this may not be the only plausible way to reason about your opponents. It often makes intuitive sense to also think critically about what your opponent has done so far, and to use his past behavior to draw some conclusions about how he may act now and in the future. To illustrate this, let us go back to the example “Painting Chris’ house”.

Example 9.1. Painting Chris’ house.

Recall the story from Example 8.1 in the previous chapter. For convenience, we have reproduced the graphical representation of the game in Figure 9.1. Let us first repeat how the concept of common belief in future rationality argues in this game. Suppose you observe that Barbara has rejected her colleague’s offer. If you believe that Barbara chooses rationally afterwards, then you believe that she will certainly not choose a price of 500, as it is strictly dominated for her by the randomized choice in which she choose the prices 200 and 400 with probability 0.5. However, if you believe that Barbara does not choose a price of 500, then

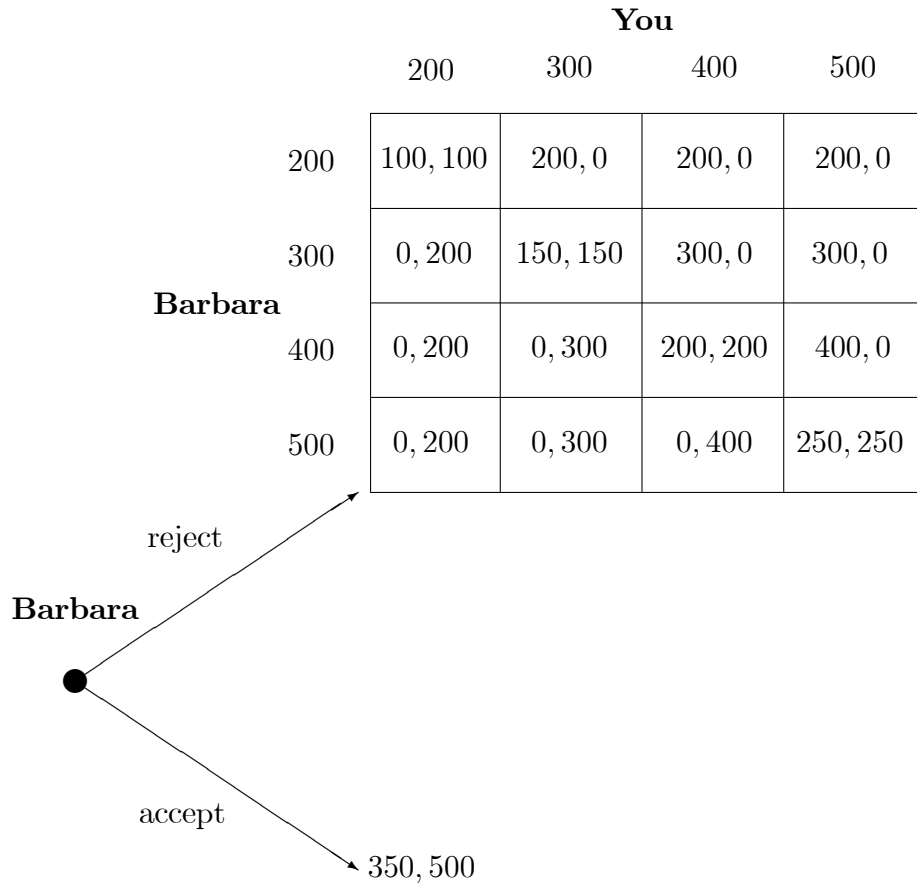
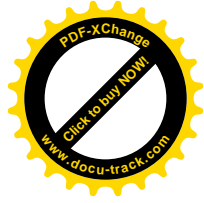
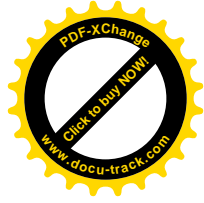


Figure 9.1: Painting Chris' house

400 and 500 can no longer be optimal prices for you. Hence, if Barbara believes that you choose rationally, and that you believe in Barbara's future rationality, then Barbara believes that you will not choose prices 400 and 500. But then, only a price of 200 can be optimal for Barbara after rejecting the colleague's offer. That is, if you believe in Barbara's future rationality, believe that Barbara believes in your rationality, and believe that Barbara believes that you believe in her future rationality, then you believe, upon observing that Barbara has rejected her colleague's offer, that Barbara will choose a price of 200. Therefore, you will choose a price of 200 as well. As such, your only rational strategy under common belief in future rationality is to choose a price of 200.



Consequently, under common belief in future rationality you are led to conclude that Barbara has made an *irrational* choice upon observing that she has rejected her colleague's offer! Under common belief in future rationality, namely, you believe that Barbara believes that you will choose a price of 200. But then, you believe that Barbara would do best by accepting her colleague's offer, so if you observe that she has rejected the offer you conclude that she must have made a mistake.

You could also have reasoned differently here. Upon observing that Barbara has rejected the colleague's offer, you could alternatively ask whether this decision can be part of a *rational* strategy for Barbara, that is, a strategy that is also optimal at the beginning of the game for some belief about your price choice. If so, then you believe, upon observing that she has rejected the offer, that she is indeed using such a rational strategy that includes rejecting the colleague's offer as a rational choice. We say that you *strongly believe in Barbara's rationality*.

Let us investigate the consequences of *strong belief in Barbara's rationality* for you. For Barbara, rejecting the colleague's offer can only be part of a rational strategy if she chooses a price of 400 afterwards. Only then, namely, will she have the chance of receiving more than the 350 euros she gets by accepting the colleague's offer. So, if you strongly believe in Barbara's rationality, then you conclude, after observing that Barbara has rejected the offer, that she will continue by choosing a price of 400, as this is the only way to turn it into a rational strategy. But then, you will choose a price of 300, and not 200 as you would do under common belief in future rationality! So, under strong belief in Barbara's rationality you would choose differently than under common belief in future rationality.

Again, it is difficult to say which of these two lines of reasoning is more appealing – they are simply different, and in our opinion both make a lot of intuitive sense. So, we present both patterns of reasoning next to each other in this book, and it is up to the reader to decide to which line of reasoning he or she feels more attracted. \square

In the example above we have described the idea of *strongly believing in Barbara's rationality* – it states that *if* it is possible for you to believe that rejecting the colleague's offer is part of a rational strategy for Barbara, then upon observing that she did reject the offer, you *must* believe that Barbara is choosing a rational strategy that includes rejecting the offer as a rational choice. We can generalize this idea to arbitrary dynamic games in the following way: Suppose that player i finds himself at an information set $h \in H_i$, and that he reasons about the opponents' strategy choices. We say that player i *strongly believes in the opponents'*



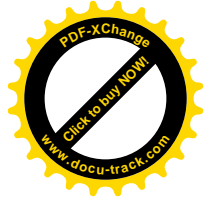
rationality at information set h if, *whenever there is* a combination of opponents' strategies that leads to h and that is optimal at every opponents' information set, player i *must* at h *only assign positive probability* to such optimal opponents' strategies. That is, if it is possible for player i to believe at h that each of his opponents chooses a rational strategy, he *must* believe at h that every opponent is choosing a rational strategy.

Note that it may not always be possible for a player to believe, at a given information set, that his opponents are using rational strategies. Consider, for instance, the example "Painting Chris' house", and suppose that Barbara could earn 450 euros (instead of 350) by accepting the colleague's offer. Then, upon observing that Barbara has rejected the colleague's offer, it is no longer possible for you to believe that Barbara is using a rational strategy. In that case, strong belief in Barbara's rationality would not impose any restrictions on what you believe after observing that Barbara has rejected the offer. In general, if it is not possible for player i to believe at information set $h \in H_i$ that each of his opponents is using a rational strategy, then the concept of strong belief in the opponents' rationality does not restrict player i 's belief at information set h at all.

Let us now see how we can formalize the idea of strong belief in the opponents' rationality, which we have described somewhat informally above. So, let us model the players' conditional belief hierarchies formally by means of an epistemic model M , in which T_i is the set of types for player i , and $b_i(t_i, h)$ specifies t_i 's conditional belief at information set $h \in H_i$ about the opponents' strategy-type combinations. How can we formalize within the epistemic model M the requirement that, *if* it is possible for player i to believe at h that every opponent chooses a rational strategy, then player i *must* at h only assign positive probability to rational opponents' strategies? As a first attempt, we could try the following condition: Fix a type t_i within the epistemic model M , and fix an information set $h \in H_i$. Say that type t_i strongly believes in the opponents' rationality at h if

- *whenever* there is an opponents' strategy-type combination in M where (a) the opponents' strategy combination leads to h , and (b) for every opponent j , the strategy is optimal for the type at every $h' \in H_j$ that the strategy leads to, then
- type t_i *must* at h only assign positive probability to strategy-type combinations in M that satisfy conditions (a) and (b) above.

However, this condition is not enough to correctly model the idea of strong belief in the opponents' rationality. Let us return, namely, to the



Types	$T_1 = \{t_1\}, T_2 = \{t_2\}$
Beliefs for Barbara	$b_1(t_1, \emptyset) = (200, t_2)$ $b_1(t_1, h_1) = (200, t_2)$
Beliefs for you	$b_2(t_2, h_1) = ((reject, 200), t_1)$

Table 9.1: An epistemic model for “Painting Chris’ house”

example “Painting Chris’ house” with the game in Figure 9.1, and consider the epistemic model M in Table 9.1. Here, h_1 is the information set after Barbara rejects the colleague’s offer. Barbara is player 1, whereas you are player 2. Within the epistemic model M , there is no type for Barbara for which rejecting the offer is part of an optimal strategy. The only type t_1 for Barbara in M , namely, believes that you will choose a price of 200, and hence rejecting the offer is not optimal for t_1 . So, the condition we attempt above would not restrict your belief at information h_1 within the epistemic model M .

However, it is clear that there *is* a type for Barbara, outside M , for which rejecting the offer is part of an optimal strategy – namely any type for Barbara that assigns a large enough probability to you choosing a price of 500. So, strong belief in the opponents’ rationality concludes in this case that you *must* believe at h_1 that Barbara is of such a type for which rejecting the offer is part of an optimal strategy. The problem, however, is that such a type for Barbara is not included in the epistemic model we consider, and this prevents us from correctly implementing the idea of strong belief in the opponents’ rationality here.

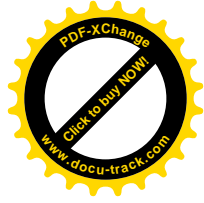
In order to solve this problem we must require, in addition, that the epistemic model contains “sufficiently many types” for the players. That is, we must add the following condition to the one above:

- If at information set $h \in H_i$ we can find a combination of opponents’ types, possibly outside M , for which there is a combination of optimal strategies leading to h , then M must contain at least one such combination of opponents’ types.

Together with the condition above, we arrive at the following formal definition of *strong belief in the opponents’ rationality*.

DEFINITION 9.1.1. (*Strong belief in the opponents’ rationality*)

Consider an epistemic model M , a type t_i for player i within M , and an information set $h \in H_i$. Type t_i **strongly believes in the opponents’**



rationality at h if, whenever we can find a combination of opponents' types, possibly outside M , for which there is a combination of optimal strategies leading to h , then

(1) the epistemic model M must contain at least one such combination of opponents' types, and

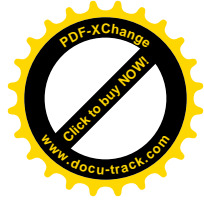
(2) type t_i must at h only assign positive probability to opponents' strategy-type combinations where the strategy combination leads to h , and the strategies are optimal for the types.

Finally, type t_i is said to strongly believe in the opponents' rationality if he does so at every information set $h \in H_i$.

Here, whenever we say that a strategy s_j is optimal for a type t_j , we precisely mean that s_j is optimal for t_j at every information set $h \in H_j$ that s_j leads to. Note that there is a remarkable similarity between this definition and the one we gave for *assuming the opponent's rationality* in Chapter 7. In both definitions we require the epistemic model to contain sufficiently many types – condition (1) in both definitions – otherwise the definitions “do not work” as we have seen.

With this formal definition at hand it can now easily be verified that in the epistemic model of Table 9.1, your type t_2 does not strongly believe in Barbara's rationality. We have seen that for Barbara we can find a belief hierarchy for which there is an optimal strategy leading to h_1 , namely any belief hierarchy that assigns large enough probability to you choosing a price of 500. However, the epistemic model M does not contain any type for Barbara for which there is an optimal strategy leading to h_1 , and hence condition (1) in the definition of strong belief in the opponents' rationality is violated.

Consider now the new epistemic model for “Painting Chris' house” in Table 9.2. We will verify that within this new epistemic model M , your type t_2 strongly believes in Barbara's rationality. Note that for Barbara's type t_1^r there is an optimal strategy leading to h_1 , namely the strategy $(reject, 400)$. Since this type t_1^r is contained in M , condition (1) in the definition of strong belief in Barbara's rationality is satisfied. Your type t_2 assigns at h_1 probability 1 to Barbara's strategy-type pair $((reject, 400), t_1^r)$, where the strategy $(reject, 400)$ leads to h_1 , and is optimal for type t_1^r . Hence, your type t_2 also satisfies condition (2) in the definition, and therefore we may conclude that your type t_2 strongly believes in Barbara's rationality.



Types	$T_1 = \{t_1^a, t_1^r\}, T_2 = \{t_2\}$
Beliefs for Barbara	$b_1(t_1^a, \emptyset) = (300, t_2)$
	$b_1(t_1^a, h_1) = (300, t_2)$
	$b_1(t_1^r, \emptyset) = (500, t_2)$
	$b_1(t_1^r, h_1) = (500, t_2)$
Beliefs for you	$b_2(t_2, h_1) = ((reject, 400), t_1^r)$

Table 9.2: A new epistemic model for “Painting Chris’ house”

9.2. Common Strong Belief in Rationality

So far we have introduced the idea of strongly believing in the opponents’ rationality, which means that, whenever possible, you believe that your opponents are implementing rational strategies. We can carry this argument one step further, however. Consider a dynamic game with two players, say i and j . Suppose that player i is at information set $h \in H_i$, and he is considering two strategies, s_j and s'_j , for opponent j that both lead to h . Assume that both strategies are optimal for some types of player j . However, s_j is also optimal for a player j type that strongly believes in i ’s rationality, whereas s'_j is only optimal for player j types that do not strongly believe in i ’s rationality. Then, intuitively, s_j is a “more plausible” strategy for player j than s'_j , as it can be supported by a “more plausible” belief hierarchy. Consequently, player i must at information set h assign probability 0 to the “less plausible” strategy s'_j as there is a “more plausible” strategy for player j leading to h as well. We say that player i expresses *2-fold strong belief in rationality*. Before attempting to formalize this requirement, let us first apply this idea to the following example.

Example 9.2. Watching TV with Barbara.

It is Wednesday evening, and Barbara and you must decide which TV program to watch this evening. The only interesting programs on TV tonight are *Blackadder* and *Dallas*. The problem is that you clearly prefer *Blackadder*, whereas Barbara wants to see *Dallas*. More precisely, watching *Blackadder* gives you a utility of 6 and Barbara a utility of 3, whereas for *Dallas* it is the other way around. As a possible resolution to this problem, you both simultaneously write down a program on a piece

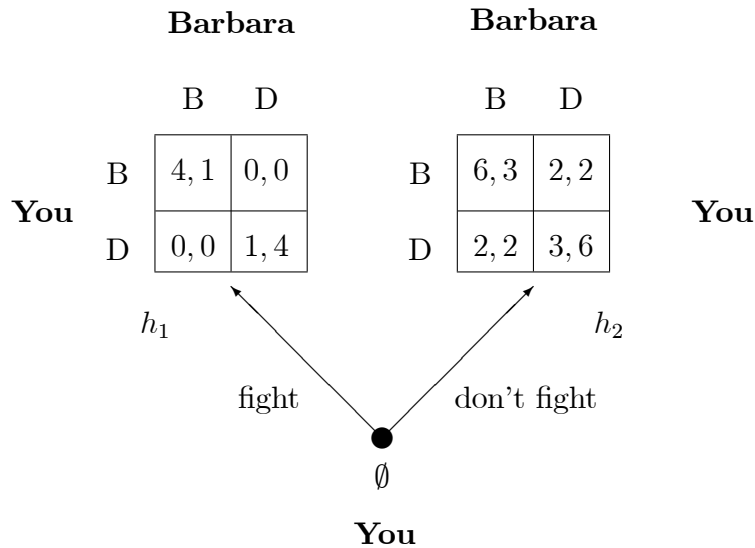
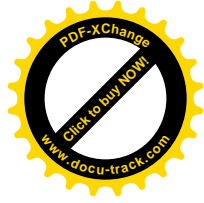


Figure 9.2: Watching TV with Barbara

of paper, and compare these. If you both write down the same program, you will watch it together. If you write down different programs, then the TV will remain switched off, and you will both play a game of cards. In that case, the utility will be only 2 for both. However, before writing down the program on a piece of paper, you have the option to start a fight with Barbara in order to convince her to watch your favorite program – something that you have done repeatedly in the past, so Barbara is well aware of this option. If you start a fight, then this would reduce both your utility and Barbara's utility by 2. This situation can be modeled by the dynamic game in Figure 9.2. Here, B stands for *Blackadder* and D stands for *Dallas*.

Suppose that the game reaches h_1 , that is, you have chosen to start a fight with Barbara. Among the two strategies for you that reach h_1 , which are $(fight, B)$ and $(fight, D)$, only strategy $(fight, B)$ can be optimal at the beginning of the game, since $(fight, D)$ is always worse than not starting a fight at the beginning of the game. So, if Barbara strongly believes in your rationality, then at h_1 she must believe that you have chosen the strategy $(fight, B)$. Hence, if Barbara strongly believes in your rationality, then she can only rationally choose B at h_1 . Consequently, under strong belief in your rationality, Barbara can only rationally choose the strategies (B, B) and (B, D) , where (B, D) means that she chooses B at h_1 and D at h_2 .



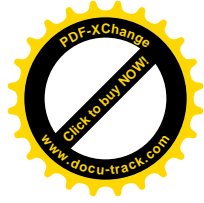
Now, what would you believe about Barbara's strategy choice in this game? Notice first that each of Barbara's strategies is optimal for some conditional belief hierarchy. That is, if you are only required to strongly believe in Barbara's rationality, then you are free to hold any belief about Barbara's strategy choice. However, we have just seen that Barbara's strategies (B, B) and (B, D) can both be supported by types for Barbara that strongly believe in your rationality, whereas her strategies (D, B) and (D, D) cannot. So, in a sense, her strategies (B, B) and (B, D) are "more plausible" than (D, B) and (D, D) . But then, if you express 2-fold strong belief in rationality, you can only assign positive probability to Barbara's strategies (B, B) and (B, D) . That is, under 2-fold strong belief in rationality you believe that Barbara will write down *Blackadder* after starting a fight with her. As a consequence, you expect a utility of 4 by starting a fight with her, and choosing *Blackadder* afterwards. Hence, under 2-fold strong belief in rationality it is no longer optimal to choose strategy $(don't, D)$, as it would yield you at most a utility of 3. Summarizing, we see that under 2-fold strong belief in rationality you can only rationally choose the strategies $(fight, B)$ and $(don't, B)$. \square

Let us now try to formalize the notion of 2-fold strong belief in rationality within an epistemic model. Consider an arbitrary dynamic game, and suppose player i finds himself at information set $h \in H_i$. The idea is that player i , if possible, only assigns positive probability to those opponents' strategy combinations that lead to h , and where every strategy can be supported by a type that strongly believes in his opponents' rationality. In order to formalize this idea correctly, we must again make sure that the epistemic model contains sufficiently many types. More precisely, if at information set h we can indeed find some combination of opponents' strategies and types, possibly outside the model, where (a) the strategy combination leads to h , (b) every strategy is optimal for the corresponding type, and (c) every type strongly believes in the opponents' rationality, then the epistemic model must contain such a combination of opponents' types.

DEFINITION 9.2.1. (*2-fold strong belief in rationality*)

Consider an epistemic model M , a type t_i for player i within M , and an information set $h \in H_i$. Type t_i expresses **2-fold strong belief in rationality** at h if, whenever we can find a combination of opponents' types, possibly outside M , that strongly believe in their opponents' rationality, and for which there is a combination of optimal strategies leading to h , then

(1) the epistemic model M must contain at least one such combination



Types	$T_1 = \{t_1^{fB}, t_1^{dB}, t_1^{dD}\}, T_2 = \{t_2^{BB}, t_2^{BD}, t_2^{DD}\}$
Beliefs for you	$b_1(t_1^{fB}) = ((B, D), t_2^{BD})$ $b_1(t_1^{dB}) = ((B, B), t_2^{BB})$ $b_1(t_1^{dD}) = ((D, D), t_2^{DD})$
Beliefs for Barbara	$b_2(t_2^{BB}, h_1) = ((fight, B), t_1^{fB})$ $b_2(t_2^{BB}, h_2) = ((don't, B), t_1^{dB})$ $b_2(t_2^{BD}, h_1) = ((fight, B), t_1^{fB})$ $b_2(t_2^{BD}, h_2) = ((don't, D), t_1^{dD})$ $b_2(t_2^{DD}, h_1) = ((fight, D), t_1^{fB})$ $b_2(t_2^{DD}, h_2) = ((don't, D), t_1^{dD})$

Table 9.3: An epistemic model for “Watching TV with Barbara”

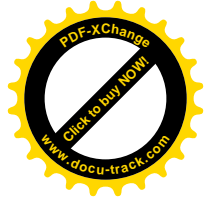
of opponents' types, and

(2) type t_i must at h only assign positive probability to opponents' strategy-type combinations where the strategy combination leads to h , the types strongly believe in their opponents' rationality, and the strategies are optimal for the types.

Finally, type t_i is said to express 2-fold strong belief in rationality if he does so at every information set $h \in H_i$.

To illustrate this definition, let us return to the example “Watching TV with Barbara”, and consider the epistemic model in Table 9.3. In that epistemic model we have not explicitly specified the information set at which your types hold the conditional belief. We have done so because we assume that your types hold the same conditional belief at \emptyset, h_1 and h_2 . Indeed, there is no reason for you to change your initial belief when the game reaches h_1 or h_2 , because Barbara does not choose before h_1 and h_2 .

We will show that your types t_1^{fB} and t_1^{dB} both express 2-fold strong belief in rationality. To do so, we first verify that Barbara's types t_2^{BB} and t_2^{BD} strongly believe in your rationality. Barbara's types t_2^{BB} and t_2^{BD} assign, at h_1 and h_2 , probability one to a strategy-type pair for you



where the strategy is optimal for the type. Therefore, we can immediately conclude that Barbara's types t_2^{BB} and t_2^{BD} strongly believe in your rationality.

Your type t_1^{fB} assigns probability 1 to Barbara's strategy-type pair $((B, D), t_2^{BD})$. Since, as we have seen, Barbara's type t_2^{BD} strongly believes in your rationality, and strategy (B, D) is optimal (at h_1 and h_2) for type t_2^{BD} , it follows that your type t_1^{fB} expresses 2-fold strong belief in rationality. In a similar fashion, you may verify that also your type t_1^{dB} expresses 2-fold strong belief in rationality.

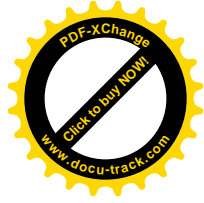
However, your type t_1^{dD} does not express 2-fold strong belief in rationality. To show this, we first verify that Barbara's type t_2^{DD} does not strongly believe in your rationality. At h_1 , there is a strategy-type pair for you where the strategy leads to h_1 and the strategy is optimal for the type, namely the strategy-type pair $((fight, B), t_1^{fB})$. Hence, for Barbara's type t_2^{DD} to strongly believe in your rationality, it must at h_1 assign positive probability only to strategy-type pairs for you where the strategy is optimal for the type. But this is not what t_2^{DD} does. At h_1 , type t_2^{DD} assigns probability 1 to your strategy-type pair $((fight, D), t_1^{fB})$, where $(fight, D)$ is not optimal for t_1^{fB} at the beginning of the game. Hence, Barbara's type t_2^{DD} does not strongly believe in your rationality.

Your type t_1^{dD} always assigns probability 1 to Barbara's type t_2^{DD} which, as we have seen, does not strongly believe in your rationality. This means, however, that your type t_1^{dD} does not express 2-fold strong belief in rationality.

Hence, the only types for you in the epistemic model that express 2-fold strong belief in rationality are t_1^{fB} and t_1^{dB} . The optimal strategies for these two types are $(fight, B)$ and $(don't, B)$ which, as we have seen above, are the only strategies that you can rationally choose under 2-fold strong belief in rationality.

At first sight, one might be tempted to conclude that if a type expresses 2-fold strong belief in rationality, then it also automatically strongly believes in the opponent's rationality. This, however, is wrong! There are types that express 2-fold strong belief in rationality, but that do not strongly believe in the opponents' rationality! As an illustration, let us go back to the example "Painting Chris' house" with the epistemic model in Table 9.1. We have already seen above that your type t_2 does not strongly believe in Barbara's rationality. We will show, however, that this same type t_2 of yours expresses 2-fold strong belief in rationality.

Notice first that there is no type for Barbara that strongly believes in your rationality, and for which rejecting the colleague's offer is optimal. Namely, if Barbara strongly believes in your rationality, then she



believes that you will not choose a price of 500, and hence she would never reject the colleague's offer. So, in the definition of 2-fold strong belief in rationality, there is no type for Barbara that strongly believes in your rationality, and for which there is an optimal strategy leading to information set h_1 . But then, 2-fold strong belief in rationality does not impose any restrictions on the beliefs you may hold at information set h_1 , which means in particular that your type t_2 expresses 2-fold strong belief in rationality. That is, your type t_2 expresses 2-fold strong belief in rationality, but does not strongly believe in Barbara's rationality.

Now that we have established a formal definition for 2-fold strong belief in rationality, we can easily extend this idea to formally define k -fold strong belief in rationality for every $k \geq 3$. The way we build up this definition is very similar to how we have defined k -fold assumption of rationality for lexicographic beliefs in Chapter 7. Namely, we first define 1-fold strong belief in rationality – which will simply be strong belief in the opponents' rationality – and will then inductively define k -fold strong belief in rationality for every $k \geq 2$. Here, for defining k -fold strong belief in rationality we will always use the definitions already obtained for 1-fold until $(k - 1)$ -fold strong belief in rationality – like we did for k -fold assumption of rationality.

DEFINITION 9.2.2. (k -fold strong belief in rationality)

Consider an epistemic model M and a type t_i for player i within M . Type t_i is said to express 1-fold strong belief in rationality if t_i strongly believes in the opponents' rationality.

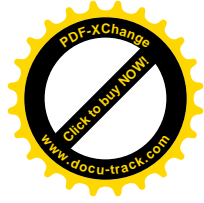
Now, fix a $k \geq 2$ and an information set $h \in H_i$. Say that type t_i expresses **k -fold strong belief in rationality** at h if, whenever we can find a combination of opponents' types, possibly outside M , that express up to $(k - 1)$ -fold strong belief in rationality, and for which there is a combination of optimal strategies leading to h , then

(1) the epistemic model M must contain at least one such combination of opponents' types, and

(2) type t_i must at h only assign positive probability to opponents' strategy-type combinations where the strategy combination leads to h , the types express up to $(k - 1)$ -fold strong belief in rationality, and the strategies are optimal for the types.

Finally, type t_i is said to express k -fold strong belief in rationality if he does so at every information set $h \in H_i$.

Here, whenever we say that a type expresses “up to $(k - 1)$ -fold strong belief in rationality”, we mean that it expresses 1-fold strong belief



in rationality, 2-fold strong belief in rationality, ... , until $(k - 1)$ -fold strong belief in rationality. Note that if we choose $k = 2$, then we obtain precisely the definition of 2-fold strong belief in rationality as provided before.

We have seen above that a type that expresses 2-fold strong belief in rationality need not express 1-fold strong belief in rationality. Hence, in general, a type that expresses k -fold strong belief in rationality need not express $(k - 1)$ -fold strong belief in rationality – a phenomenon we have already seen for k -fold assumption of rationality in Chapter 7.

With the definition of k -fold strong belief in rationality at our disposal, it is now easy to define *common* strong belief in rationality – it simply means that the type expresses k -fold strong belief in rationality for every k .

DEFINITION 9.2.3. (*Common strong belief in rationality*)

Consider an epistemic model M and a type t_i for player i within M . Type t_i is said to express **common strong belief in rationality** if t_i expresses k -fold strong belief in rationality for every k .

Player i can rationally choose strategy s_i under common strong belief in rationality if there is some epistemic model M and some type t_i within M such that

- (1) type t_i expresses common strong belief in rationality, and
- (2) strategy s_i is optimal for type t_i at every information set $h \in H_i$ that s_i leads to.

Let us illustrate this definition by means of the example “Watching TV with Barbara”.

Example 9.3. Watching TV with Barbara.

Consider the graphical representation in Figure 9.2 and the epistemic model in Table 9.3. We have seen above that your types t_1^{fB} and t_1^{dB} express 2-fold strong belief in rationality but that your type t_1^{dD} does not. Your type t_1^{dD} does express 1-fold strong belief in rationality, however. Moreover, we have seen that Barbara’s types t_2^{BB} and t_2^{BD} express 1-fold strong belief in rationality, but that her type t_2^{DD} does not. We will now show that your types t_1^{fB} and t_1^{dB} also express 1-fold strong belief in rationality, and that Barbara’s types t_2^{BB} and t_2^{BD} also express 2-fold strong belief in rationality.

Your types t_1^{fB} and t_1^{dB} both assign, at each information set, probability 1 to a strategy-type pair for Barbara where the strategy is optimal



for the type. This is enough to conclude that these two types express 1-fold strong belief in rationality.

Let us now turn to Barbara's type t_2^{BB} . At information set h_1 (if you decide to fight) there is a type for you that expresses 1-fold strong belief in rationality, and an optimal strategy for that type leading to h_1 , namely your type t_1^{fB} with optimal strategy $(fight, B)$. So, in order for t_2^{BB} to express 2-fold strong belief in rationality at h_1 , it must at h_1 assign positive probability only to strategy-type pairs for you where the type expresses 1-fold strong belief in rationality, and the strategy is optimal for the type. But that is what t_2^{BB} does as h_1 . Namely, type t_2^{BB} assigns at h_1 probability 1 to the strategy-type pair $((fight, B), t_1^{fB})$, where t_1^{fB} expresses 1-fold strong belief in rationality, and $(fight, B)$ is optimal for t_1^{fB} .

At information set h_2 (if you decide not to fight) there is also a type for you that expresses 1-fold strong belief in rationality, and an optimal strategy for that type leading to h_2 , namely your type t_1^{dB} with optimal strategy $(don't, B)$, or your type t_1^{dD} with optimal strategy $(don't, D)$. So, in order for t_2^{BB} to express 2-fold strong belief in rationality at h_2 , it must at h_2 assign positive probability only to strategy-type pairs for you where the type expresses 1-fold strong belief in rationality, and the strategy is optimal for the type. But that is what t_2^{BB} does at h_2 . Namely, type t_2^{BB} assigns at h_2 probability 1 to the strategy-type pair $((don't, B), t_1^{dB})$, where t_1^{dB} expresses 1-fold strong belief in rationality, and $(don't, B)$ is optimal for t_1^{dB} . Hence, we may conclude that Barbara's type t_2^{BB} expresses 2-fold strong belief in rationality at h_1 and h_2 , and therefore expresses 2-fold strong belief in rationality overall.

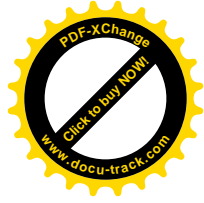
In the same way, it can be verified that also Barbara's type t_2^{BD} expresses 2-fold strong belief in rationality.

So, overall we see that your types t_1^{fB} and t_1^{dB} , and Barbara's types t_2^{BB} and t_2^{BD} express up to 2-fold strong belief in rationality, whereas the other types t_1^{dD} and t_2^{DD} do not.

We will now show that your types t_1^{fB} and t_1^{dB} also express 3-fold strong belief in rationality.

Let us first turn to your type t_1^{fB} . At every information set, your type t_1^{fB} assigns probability 1 to Barbara's strategy-type pair $((B, D), t_2^{BD})$, where t_2^{BD} expresses up to 2-fold strong belief in rationality, and (B, D) is optimal for t_2^{BD} . This is enough to conclude that your type t_1^{fB} expresses 3-fold strong belief in rationality. In a similar way, it can be checked that also your type t_1^{dB} expresses 3-fold strong belief in rationality.

Next, we show that Barbara's type t_2^{BB} also expresses 3-fold strong belief in rationality. By definition, Barbara's type t_2^{BB} assigns at h_1



probability 1 to your strategy-type pair $((fight, B), t_1^{fB})$, and assigns at h_2 probability 1 to your strategy-type pair $((don't, B), t_1^{dB})$. Since t_1^{fB} and t_1^{dB} express up to 2-fold strong belief in rationality, strategy $(fight, B)$ is optimal for t_1^{fB} , and strategy $(don't, B)$ is optimal for t_1^{dB} , we may immediately conclude that Barbara's type t_2^{BB} expresses 3-fold strong belief in rationality.

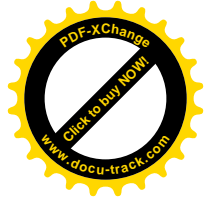
We show, however, that Barbara's type t_2^{BD} does not express 3-fold strong belief in rationality. At h_2 there is a strategy-type pair for you where the strategy leads to h_2 , the type expresses up to 2-fold strong belief in rationality, and the strategy is optimal for the type, namely the strategy-type pair $((don't, B), t_1^{dB})$. So, for t_2^{BD} to express 3-fold strong belief in rationality, it must assign at h_2 only positive probability to strategy-type pairs for you where the strategy is optimal for the type and the type expresses up to 2-fold strong belief in rationality. This, however, is not what t_2^{BD} does. At h_2 , type t_2^{BD} assigns probability 1 to your strategy-type pair $((don't, D), t_1^{dD})$, but we have seen that t_1^{dD} does not express up to 2-fold strong belief in rationality. Hence, Barbara's type t_2^{BD} does not express 3-fold strong belief in rationality.

Summarizing so far, we see that your types t_1^{fB} and t_1^{dB} , and Barbara's type t_2^{BB} , express up to 3-fold strong belief in rationality, whereas the other types do not.

We now prove that your type t_1^{dB} and Barbara's type t_2^{BB} also express 4-fold strong belief in rationality, but that your type t_1^{fB} does not.

For your type t_1^{dB} this is easily seen, as t_1^{dB} always assigns probability 1 to Barbara's strategy-type pair $((B, B), t_2^{BB})$, where t_2^{BB} expresses up to 3-fold strong belief in rationality, and strategy (B, B) is optimal for type t_2^{BB} . This is enough to conclude that your type t_1^{dB} expresses 4-fold strong belief in rationality.

Consider now your type t_1^{fB} . Clearly, at each of your information sets there is a strategy-type pair for Barbara where the strategy leads to that information set, the strategy is optimal for the type, and the type expresses up to 3-fold strong belief in rationality, namely the strategy-type pair $((B, B), t_2^{BB})$. So, in order for t_1^{fB} to express 4-fold strong belief in rationality, it must at every information set only assign positive probability to strategy-type pairs for Barbara where the strategy is optimal for the type and the type expresses up to 3-fold strong belief in rationality. But this is not what t_1^{fB} does. Type t_1^{fB} assigns probability 1 to Barbara's type t_2^{BD} which, as we have seen, does not express up to 3-fold strong belief in rationality. Hence, your type t_1^{fB} does not express 4-fold strong belief in rationality.



We will now show that Barbara's type t_2^{BB} does express 4-fold strong belief in rationality. Note that type t_2^{BB} assigns at h_1 probability 1 to your strategy-type pair $((fight, B), t_1^{fB})$ and assigns at h_2 probability 1 to your strategy-type pair $((don't, B), t_1^{dB})$. We know, from before, that $(fight, B)$ is optimal for t_1^{fB} , that $(don't, B)$ is optimal for t_1^{dB} , and that the types t_1^{fB} and t_1^{dB} both express up to 3-fold strong belief in rationality. This is enough to conclude that Barbara's type t_2^{BB} expresses 4-fold strong belief in rationality.

Summarizing, we see that your type t_1^{dB} and Barbara's type t_2^{BB} express up to 4-fold strong belief in rationality, and that the other types do not.

We finally show that your type t_1^{dB} and Barbara's type t_2^{BB} also express 5-fold strong belief in rationality and further.

Let us start with your type t_1^{dB} and see why it expresses 5-fold strong belief in rationality. This is easily seen, in fact, as t_1^{dB} always assigns probability 1 to Barbara's strategy-type pair $((B, B), t_2^{BB})$, where (B, B) is optimal for t_2^{BB} and, as we have seen, type t_2^{BB} expresses up to 4-fold strong belief in rationality.

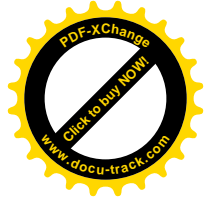
To verify the same for Barbara's type t_2^{BB} is more difficult, however. Consider the information set h_1 , where you have decided to start a fight with Barbara. We show that there is *no* type for you which expresses up to 4-fold strong belief in rationality, and that has an optimal strategy leading to h_1 .

We have seen that only your strategies $(fight, B)$, $(don't, B)$ and $(don't, D)$ are optimal at the beginning for some belief about Barbara's strategy choice. Hence, under 1-fold strong belief in rationality Barbara must conclude at h_1 that you are choosing $(fight, B)$. So, the only optimal strategies for Barbara under 1-fold strong belief in rationality are (B, B) and (B, D) .

Consequently, under 2-fold strong belief in rationality you must believe, throughout the game, that Barbara either chooses (B, B) or (B, D) . As such, the only optimal strategies for you under expressing up to 2-fold strong belief in rationality are $(fight, B)$ and $(don't, B)$.

But then, under 3-fold strong belief in rationality Barbara must conclude at h_1 that you are choosing $(fight, B)$, and must conclude at h_2 that you are choosing $(don't, B)$. Hence, the only optimal strategy for Barbara if she expresses up to 3-fold strong belief in rationality is (B, B) .

So, under 4-fold strong belief in rationality you must believe that Barbara chooses (B, B) . This means that there is only one strategy you can rationally choose under expressing up to 4-fold strong belief in rationality, namely $(don't, B)$. Hence, indeed, there is no type for you that



expresses up to 4-fold strong belief in rationality, and that has an optimal strategy leading to h_1 .

But then, 5-fold strong belief in rationality does not impose any restrictions on Barbara's belief at h_1 . In particular, Barbara's type t_2^{BB} expresses 5-fold strong belief in rationality at h_1 . At information set h_2 , Barbara's type t_2^{BB} assigns probability 1 to your strategy-type pair $((don't, B), t_1^{dB})$. Since type t_1^{dB} expresses up to 4-fold strong belief in rationality, as we have seen, and strategy $(don't, B)$ is optimal for type t_1^{dB} , it follows that Barbara's type t_2^{BB} expresses 5-fold strong belief in rationality at h_2 . Hence, we may conclude that t_2^{BB} expresses 5-fold strong belief in rationality overall.

Summarizing, we see that your type t_1^{dB} and Barbara's type t_2^{BB} express up to 5-fold strong belief in rationality.

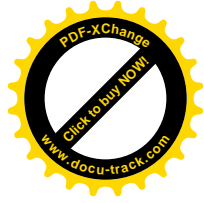
In a similar fashion, it can be verified that both types also express 6-fold strong belief in rationality and further. That is, both your type t_1^{dB} and Barbara's type t_2^{BB} express common strong belief in rationality. Since your strategy $(don't, B)$ is optimal for your type t_1^{dB} , we see that you can rationally choose strategy $(don't, B)$ under common strong belief in rationality.

We can actually show a little more: Under common strong belief in rationality, your *only* optimal strategy is $(don't, B)$! Above we have argued, namely, that if you express up to 4-fold strong belief in rationality, then your only optimal strategy is $(don't, B)$. But then, it will certainly be the only optimal strategy under *common* strong belief in rationality. Moreover, we have also seen that if you express up to 4-fold strong belief in rationality, then you believe that Barbara will choose (B, B) , that is, that Barbara will always write down *Blackadder*. But then, under common strong belief in rationality, your only possible belief is that Barbara will always write down *Blackadder*, which means that you expect to obtain the maximum utility of 6 by not starting a fight, and writing down *Blackadder* subsequently. Hence, we see that under common strong belief in rationality you expect to end up with the best possible scenario – namely that you will watch your favorite program *Blackadder* together, without having to start a fight with Barbara about this! \square

Let us finally go back to the example “Painting Chris’ house” and apply the idea of common strong belief in rationality there.

Example 9.4. Painting Chris’ house.

Reconsider the dynamic game in Figure 9.1 and the epistemic model in Table 9.2. We have already seen that your type t_2 expresses 1-fold strong



belief in rationality. We will show that your type t_2 and Barbara's type t_1^a both express common strong belief in rationality, but Barbara's type t_1^r not.

In fact, Barbara's type t_1^r does not even express 1-fold strong belief in rationality. Her type t_1^r , namely, assigns at the beginning of the game probability 1 to your strategy-type pair $(500, t_2)$. But strategy 500 is not optimal for your type t_2 , and hence Barbara's type t_1^r does not strongly believe in your rationality, that is, does not express 1-fold strong belief in rationality.

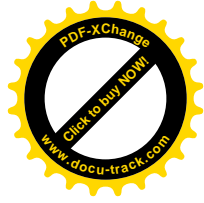
We next show that Barbara's type t_1^a expresses 1-fold strong belief in rationality. This is easily seen, as t_1^a always assigns probability 1 to your strategy-type pair $(300, t_2)$, and strategy 300 is optimal for your type t_2 . So, both your type t_2 and Barbara's type t_1^a express 1-fold strong belief in rationality.

Subsequently, we prove that your type t_2 and Barbara's type t_1^a also express 2-fold strong belief in rationality. Let us start with your type t_2 . Is there a strategy-type pair for Barbara where the type expresses 1-fold strong belief in rationality, the strategy is optimal for the type, and the strategy leads to h_1 ? The answer is "no". Namely, if Barbara expresses 1-fold strong belief in rationality – that is, strongly believes in your rationality – then she believes that you will not choose a price of 500. But then, it cannot be optimal for Barbara to reject the colleague's offer. So, there is no type for Barbara that expresses 1-fold strong belief in rationality, and which has an optimal strategy leading to h_1 . But then, 2-fold strong belief in rationality does not impose any restrictions on your beliefs at h_1 . In particular, your type t_2 expresses 2-fold strong belief in rationality.

It is easily verified that Barbara's type t_1^a expresses 2-fold strong belief in rationality: Her type t_1^a always assigns probability 1 to your strategy-type pair $(300, t_2)$, where t_2 expresses 1-fold strong belief in rationality, and strategy 300 is optimal for t_2 . So, we conclude that both your type t_2 and Barbara's type t_1^a express up to 2-fold strong belief in rationality.

In a similar fashion it can be verified that the types t_2 for you and t_1^a for Barbara also express 3-fold strong belief in rationality and further. That is, t_2 and t_1^a both express common strong belief in rationality. As strategy 300 is optimal for your type t_2 , we conclude that you can rationally choose a price of 300 under common strong belief in rationality.

In fact, 300 is the *only* price you can rationally choose under common strong belief in rationality. Namely, if you strongly believe in Barbara's rationality, then you must believe, after observing that Barbara has rejected the colleague's offer, that Barbara will choose a price of 400 afterwards, as this is the only price by which she can get more than 350



– the amount she could have obtained by accepting the colleague’s offer. But then, the only optimal price for you would be to choose 300.

So, if Barbara expresses common strong belief in rationality, then she must believe that you will choose a price of 300 in case she rejects the colleague’s offer. Therefore, under common strong belief in rationality there is only one optimal strategy for Barbara, namely to *accept* the colleague’s offer at the beginning of the game. \square

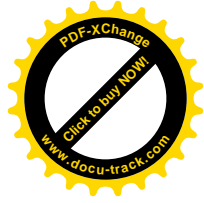
9.3. Algorithm

Up to this stage we have introduced and formalized the idea of common strong belief in rationality. We now investigate whether there is an easy algorithm that generates precisely those strategies you can rationally choose under common strong belief in rationality. It turns out that we can indeed find such an algorithm, and in fact it will be very similar to the backward dominance procedure we presented for common belief in future rationality. On the road towards this algorithm we start by characterizing those strategies you can rationally choose under 1-fold strong belief in rationality, then we characterize the strategies you can rationally choose under expressing up to 2-fold strong belief in rationality, and so on.

Step 1: 1-fold strong belief in rationality. We start by asking the following question: Which strategies can you rationally choose under 1-fold strong belief in rationality, that is, if you strongly believe in the opponents’ rationality? Remember that 1-fold strong belief in rationality means the following: If player i finds himself at information set $h \in H_i$, and if there is some opponents’ strategy combination that leads to h and where every strategy is optimal, then player i should at h only assign positive probability to such optimal opponents’ strategy combinations. If there is no such optimal opponents’ strategy combination leading to h , then no conditions are being imposed on player i ’s beliefs at h .

Here, by an “optimal strategy” we mean a strategy s_j that is optimal, at every information set $h' \in H_j$ it leads to, for *some* conditional belief there. We can rephrase the definition of 1-fold strong belief in rationality as follows: Player i should at information set h assign probability 0 to every opponent’s strategy that is not optimal, *unless* there is no optimal strategy combination that leads to h . In the latter case, no conditions are imposed on player i ’s beliefs at h .

From Chapter 8 we know that an opponent’s strategy s_j is optimal at an information set $h' \in H_j$ for *some* conditional belief there, precisely when s_j is not strictly dominated within the *full decision problem* $\Gamma^0(h')$.



Remember that the full decision problem $\Gamma^0(h') = (S_j(h'), S_{-j}(h'))$ contains for player j only the set $S_j(h')$ of strategies leading to h' , and contains for the opponents only the set $S_{-j}(h')$ of opponents' strategy combinations leading to h' . This means that an opponent's strategy s_j is not optimal exactly when it is strictly dominated within some full decision problem $\Gamma^0(h')$ at which j is active. Hence, 1-fold strong belief in rationality can alternatively be stated as follows: Player i should at information set h assign probability 0 to every strategy for opponent j that is strictly dominated at some full decision problem $\Gamma^0(h')$ at which j is active, *unless* this would rule out all possible beliefs for player i at h . In the latter case, no conditions are imposed on player i 's beliefs at h .

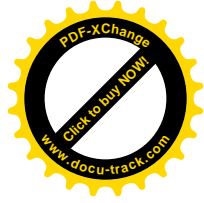
Now, requiring that player i should at h assign probability 0 to such strategies for opponent j can be mimicked by "removing" these opponent's strategies from the full decision problem $\Gamma^0(h)$ at h . But then, 1-fold strong belief in rationality can be mimicked by the following elimination step: At the full decision problem $\Gamma^0(h)$, eliminate for every opponent j those strategies that are strictly dominated within some full decision problem $\Gamma^0(h')$ at which j is active, *unless* this would remove all strategy combinations leading to h . In the latter case, we do not remove any strategies from h . Let $\Gamma^1(h)$ be the reduced decision problem at h that remains after removing strategies from $\Gamma^0(h)$ in this way.

So far we have shown that, if player i expresses 1-fold strong belief in rationality, then at every information set $h \in H_i$ he should assign positive probability only to opponents' strategy combinations in $\Gamma^1(h)$. In that case, the strategies he can rationally choose are precisely those strategies that are optimal, at every information set $h \in H_i$ it leads to, for some conditional belief at $\Gamma^1(h)$. But, as we know from Chapter 8, these are precisely the strategies that are not strictly dominated within any decision problem $\Gamma^1(h)$ at which player i is active.

We thus see that for every player i , the strategies he can rationally choose under 1-fold strong belief in rationality are precisely the strategies that are not strictly dominated at any reduced decision problem $\Gamma^1(h)$ at which he is active. In turn, these are precisely the strategies that remain after removing, from the reduced decision problem $\Gamma^1(\emptyset)$ at the beginning of the game, those strategies for player i that *are* strictly dominated at some reduced decision problem $\Gamma^1(h)$ at which i is active.

Summarizing, the strategies that can rationally be chosen under 1-fold strong belief in rationality are obtained by the following two-step elimination procedure:

First, eliminate at every full decision problem $\Gamma^0(h)$ all strategies s_j that are strictly dominated at some full decision problem $\Gamma^0(h')$ at which



player j is active, *unless* this would remove all strategy combinations leading to h . In the latter case, do not remove any strategies from $\Gamma^0(h)$. This leads to the new decision problems $\Gamma^1(h)$ at every information set h .

Second, eliminate from the decision problem $\Gamma^1(\emptyset)$ at the beginning of the game, and for every player i , those strategies that are strictly dominated at some reduced decision problem $\Gamma^1(h)$ at which i is active. This leads to the new decision problem $\Gamma^2(\emptyset)$ at the beginning of the game.

The strategies that can rationally be chosen under 1-fold strong belief in rationality are exactly the strategies in $\Gamma^2(\emptyset)$.

Step 2: Up to 2-fold strong belief in rationality. We now go one step further, and try to characterize those strategies you can rationally choose under expressing up to 2-fold strong belief in rationality. Consider some player i and an information set $h \in H_i$ at which he is active. Then, for player i to express 2-fold strong belief in rationality at h means the following: If there is an opponents' strategy combination leading to h that can rationally be chosen under 1-fold strong belief in rationality, then player i must at h only assign positive probability to such opponents' strategy combinations.

We have seen in the step above that the players' strategies that can rationally be chosen under 1-fold strong belief in rationality are precisely the strategies in $\Gamma^2(\emptyset)$. So, 2-fold strong belief in rationality at h then means the following for player i : If there is an opponents' strategy combination in $\Gamma^2(\emptyset)$ leading to h , then player i must at h only assign positive probability to opponents' strategy combinations in $\Gamma^2(\emptyset)$. Or, equivalently, we remove from the decision problem $\Gamma^1(h)$ those opponents' strategies that are not in $\Gamma^2(\emptyset)$, *unless* this would remove all strategy combinations leading to h .

By construction, a strategy s_j for player j is not in $\Gamma^2(\emptyset)$ if it is strictly dominated at some reduced decision problem $\Gamma^1(h')$ at which player j is active. So, we remove from $\Gamma^1(h)$ those strategies s_j that are strictly dominated at some reduced decision problem $\Gamma^1(h')$ at which player j is active, *unless* this would remove all strategy combinations leading to h . In the latter case, we do not remove any further strategies from $\Gamma^1(h)$. By doing so at every h , we obtain newly reduced decision problems $\Gamma^2(h)$ at every information set h . Hence, a player who expresses up to 2-fold strong belief in rationality assigns at each of his information sets h only positive probability to opponents' strategy combinations in $\Gamma^2(h)$.

But then, the strategies that player i can rationally choose under expressing up to 2-fold strong belief in rationality will be precisely those



strategies that are optimal, at every information set $h \in H_i$ it leads to, for some conditional belief in $\Gamma^2(h)$. These, in turn, are exactly the strategies s_i that are not strictly dominated within any reduced decision problem $\Gamma^2(h)$ it leads to.

So, the strategies that can rationally be chosen under expressing up to 2-fold strong belief in rationality are obtained by the following procedure:

First, eliminate at every reduced decision problem $\Gamma^1(h)$ all strategies s_j that are strictly dominated at some reduced decision problem $\Gamma^1(h')$ at which player j is active, *unless* this would remove all strategy combinations leading to h . In the latter case, do not remove any further strategies from $\Gamma^1(h)$. This leads to the newly reduced decision problems $\Gamma^2(h)$ at every information set h .

Second, eliminate from the decision problem $\Gamma^2(\emptyset)$ at the beginning of the game, and for every player i , those strategies that are strictly dominated at some reduced decision problem $\Gamma^2(h)$ at which i is active. This leads to the new decision problem $\Gamma^3(\emptyset)$ at the beginning of the game.

The strategies that can rationally be chosen under expressing up to 2-fold strong belief in rationality are exactly the strategies in $\Gamma^3(\emptyset)$.

By repeating this argument, we can construct for every $k \geq 3$ and every information set h a reduced decision problem $\Gamma^k(h)$, and conclude that $\Gamma^k(\emptyset)$ contains exactly those strategies that can rationally be chosen by players who express up to $(k - 1)$ -fold strong belief in rationality. So, by iterating this reduction process until no further strategies can be removed, we obtain an algorithm that yields precisely those strategies that can rationally be chosen under common strong belief in rationality. This algorithm is called the *iterated conditional dominance procedure*, and is formally presented below.

ALGORITHM 9.3.1. (*Iterated conditional dominance procedure*)

Step 1. At every full decision problem $\Gamma^0(h)$, eliminate for every player i those strategies that are strictly dominated at some full decision problem $\Gamma^0(h')$ at which player i is active, *unless* this would remove all strategy combinations that lead to h . In the latter case, we do not remove any strategies from $\Gamma^0(h)$. This leads to reduced decision problems $\Gamma^1(h)$ at every information set h .

Step 2. At every reduced decision problem $\Gamma^1(h)$, eliminate for every player i those strategies that are strictly dominated at some reduced decision problem $\Gamma^1(h')$ at which player i is active, *unless* this would remove all strategy combinations that lead to h . In the latter case, we do not remove any strategies from $\Gamma^1(h)$. This leads to new reduced decision



problems $\Gamma^2(h)$ at every information set.

And so on. Continue until no more strategies can be eliminated in this way.

Like in Chapter 8, eliminating a strategy s_i from a full decision problem $\Gamma^0(h)$ formally means the following: If player i is active at h , and the full decision problem $\Gamma^0(h)$ is given by $(S_i(h), S_{-i}(h))$, then we simply eliminate strategy s_i from $S_i(h)$. If player j is active at h but not i , and the full decision problem $\Gamma^0(h)$ is given by $(S_j(h), S_{-j}(h))$, then we eliminate from $S_{-j}(h)$ every strategy combination that contains strategy s_i for player i . Similarly for eliminating a strategy from a *reduced* decision problem.

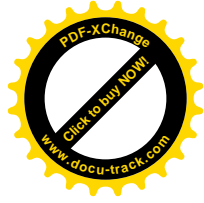
Note that this algorithm is very similar to the backward dominance procedure discussed in the previous chapter, as it also proceeds by successively eliminating strategies from decision problems at information sets. However, the criterion to eliminate a strategy at a decision problem is different: In the backward dominance procedure, we eliminate a strategy s_i in a decision problem at h if s_i is strictly dominated within a decision problem *weakly following* h at which player i is active. In the iterated conditional dominance procedure, we would also eliminate s_i if it is strictly dominated within a decision problem that comes *before* h , or that comes *neither before nor after* h – as long as this would not remove all strategy combinations leading to h .

Like for the backward dominance procedure, it can easily be seen that the iterated conditional dominance procedure will always stop within finitely many steps. Namely, there are only finitely many strategies for every player. Since the algorithm proceeds by successively eliminating strategies from decision problems at information sets, there must be a step in the procedure in which no further strategies can be eliminated, and this is where the algorithm stops.

We say that a strategy *survives* the iterated conditional dominance procedure if it is never eliminated in the decision problem at the beginning of the game.

DEFINITION 9.3.2. (*Strategy surviving the iterated conditional dominance procedure*)

For every information set h and every k , let $\Gamma^k(h)$ be the reduced decision problem produced in step k of the iterated conditional dominance procedure. Strategy s_i for player i **survives the iterated conditional dominance procedure** if s_i is in the decision problem $\Gamma^k(\emptyset)$ for every k .



It is clear that for every player i there will be at least one strategy that survives the iterated conditional dominance procedure, as we can never eliminate *all* strategies for a player at \emptyset . Our main theorem below states that the strategies that can rationally be chosen under common strong belief in rationality are exactly the strategies that survive the iterated conditional dominance procedure. That is, the algorithm we presented above gives us precisely what we want – namely the strategies that are optimal under common strong belief in rationality.

THEOREM 9.3.3. (*Algorithm “works”*)

- (1) For every $k \geq 1$, the strategies that can rationally be chosen by a type that expresses up to k -fold strong belief in rationality are precisely the strategies in $\Gamma^{k+1}(\emptyset)$ – that is, those strategies that survive step $k + 1$ of the iterated conditional dominance procedure at \emptyset .
- (2) The strategies that can rationally be chosen by a type that expresses common strong belief in rationality are exactly the strategies that survive the full iterated conditional dominance procedure at \emptyset .

The proof can be found in the proofs section at the end of this chapter. This theorem shows, in particular, that common strong belief in rationality is always possible. That is, in every dynamic game we can find for every player a type that expresses common strong belief in rationality. We have seen, namely, that the iterated conditional dominance procedure yields for every player at least one strategy. In view of the theorem, this strategy can rationally be chosen by a type that expresses common strong belief in rationality. So, in particular we can build an epistemic model in which some type of this player expresses common strong belief in rationality. In other words, common strong belief in rationality is always possible in every game. Unfortunately, we cannot provide an easy method to prove the existence, like we did for common belief in future rationality.

COROLLARY 9.3.4. (*Common strong belief in rationality is always possible*)

We can always build an epistemic model which contains, for every player i , some type t_i that expresses common strong belief in rationality.

However, for most games we cannot construct an epistemic model in which *all* types express common strong belief in rationality. Consider, for instance, the example “Painting Chris’ house”. Suppose we build an epistemic model in which there is a type t_2 for you that expresses common strong belief in rationality. In particular, this type t_2 must strongly believe in Barbara’s rationality. So, at information set h_1 –



where Barbara has decided to reject the colleague's offer – type t_2 must believe that Barbara is of a type t_1 for which it is optimal to indeed reject the offer. But, as we have seen, there is no type for Barbara that expresses common strong belief in rationality and for which rejecting the colleague's offer is optimal. That is, your type t_2 must at information set h_1 believe that Barbara is of a type t_1 which does not express common strong belief in rationality. In particular, the epistemic model at hand must contain at least one type for Barbara that does not express common strong belief in rationality. Summarizing, we see that every epistemic model for “Painting Chris' house” contains at least one type that does *not* express common strong belief in rationality.

Compare this to the concept of common belief in future rationality. In Theorem 8.7.1 of Chapter 8, we have shown that we can always construct an epistemic model in which *all* types express common belief in future rationality. This is not possible if we turn to common strong belief in rationality!

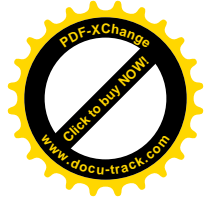
Let us now apply the iterated conditional dominance procedure to the two examples we have discussed so far in this chapter, and verify that it indeed yields precisely those strategies that can rationally be chosen under common strong belief in rationality.

Example 9.5. Painting Chris' house.

Consider the graphical representation of this game in Figure 9.1. Let h_1 be the information set that follows Barbara's choice “reject”. So, the two full decision problems $\Gamma^0(\emptyset)$ and $\Gamma^0(h_1)$ are given by Table 9.4.

Step 1. At the decision problem $\Gamma^0(\emptyset)$, the strategies $(r, 200)$, $(r, 300)$ and $(r, 500)$ for Barbara are all strictly dominated by the strategy *accept*. We therefore do not only eliminate these strategies from the decision problem $\Gamma^0(\emptyset)$, but also from the *future* decision problem $\Gamma^0(h_1)$, since by doing so we do not remove all of Barbara's strategies leading to h_1 . Note, namely, that strategy $(r, 400)$ for Barbara leads to h_1 , but is not eliminated at h_1 . This is crucially different from how the backward dominance procedure works in this example. According to the backward dominance procedure, we cannot eliminate the strategies $(r, 200)$ and $(r, 300)$ at $\Gamma^0(h_1)$, since these strategies are only strictly dominated in the decision problem $\Gamma^0(\emptyset)$ at the beginning, but not at $\Gamma^0(h_1)$.

For you, strategy 500 is strictly dominated by the randomized strategy $(0.5) \cdot 200 + (0.5) \cdot 400$ in $\Gamma^0(h_1)$. We therefore eliminate your strategy 500 from the decision problems at \emptyset and h_1 . This leads to the reduced decision problems $\Gamma^1(\emptyset)$ and $\Gamma^1(h_1)$ in Table 9.5.



$\Gamma^0(\emptyset)$: Barbara active				
	200	300	400	500
$(r, 200)$	100, 100	200, 0	200, 0	200, 0
$(r, 300)$	0, 200	150, 150	300, 0	300, 0
$(r, 400)$	0, 200	0, 300	200, 200	400, 0
$(r, 500)$	0, 200	0, 300	0, 400	250, 250
<i>accept</i>	350, 500	350, 500	350, 500	350, 500

$\Gamma^0(h_1)$: Barbara and you active				
	200	300	400	500
$(r, 200)$	100, 100	200, 0	200, 0	200, 0
$(r, 300)$	0, 200	150, 150	300, 0	300, 0
$(r, 400)$	0, 200	0, 300	200, 200	400, 0
$(r, 500)$	0, 200	0, 300	0, 400	250, 250

Table 9.4: Full decision problems in “Painting Chris’ house”

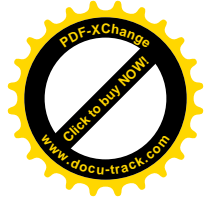
$\Gamma^1(\emptyset)$: Barbara active			
	200	300	400
$(r, 400)$	0, 200	0, 300	200, 200
<i>accept</i>	350, 500	350, 500	350, 500

$\Gamma^1(h_1)$: Barbara and you active			
	200	300	400
$(r, 400)$	0, 200	0, 300	200, 200

Table 9.5: Reduced decision problems after step 1 in “Painting Chris’ house”

Step 2. Within the reduced decision problem $\Gamma^1(\emptyset)$ at the beginning, Barbara’s strategy $(r, 400)$ is strictly dominated by *accept*. We therefore eliminate her strategy $(r, 400)$ from $\Gamma^1(\emptyset)$. However, we do not eliminate this strategy $(r, 400)$ from the future decision problem $\Gamma^1(h_1)$, since by doing so we would remove all of Barbara’s strategies leading to h_1 !

For you, strategies 200 and 400 are both strictly dominated by 300 in the reduced decision problem $\Gamma^1(h_1)$ at h_1 . We therefore eliminate your strategies 200 and 400 from $\Gamma^1(h_1)$ but also from $\Gamma^1(\emptyset)$. This leads to the final decision problems in Table 9.6, from which no further strategies can be eliminated. So, the only strategies that survive the iterated con-



$\Gamma^2(\emptyset)$: Barbara active

	300
<i>accept</i>	350, 500

$\Gamma^2(h_1)$: Barbara and you active

	300
$(r, 400)$	0, 300

Table 9.6: Final decision problems after step 2 in “Painting Chris’ house”

$\Gamma^0(\emptyset)$: You active

	(B, B)	(B, D)	(D, B)	(D, D)
$(fight, B)$	4, 1	4, 1	0, 0	0, 0
$(fight, D)$	0, 0	0, 0	1, 4	1, 4
$(don't, B)$	6, 3	2, 2	6, 3	2, 2
$(don't, D)$	2, 2	3, 6	2, 2	3, 6

$\Gamma^0(h_1)$: You and Barbara active

	(B, B)	(B, D)	(D, B)	(D, D)
$(fight, B)$	4, 1	4, 1	0, 0	0, 0
$(fight, D)$	0, 0	0, 0	1, 4	1, 4

$\Gamma^0(h_2)$: You and Barbara active

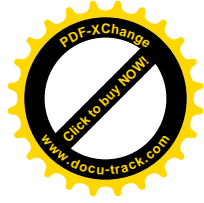
	(B, B)	(B, D)	(D, B)	(D, D)
$(don't, B)$	6, 3	2, 2	6, 3	2, 2
$(don't, D)$	2, 2	3, 6	2, 2	3, 6

Table 9.7: Full decision problems in “Watching TV with Barbara”

ditional dominance procedure are *accept* for Barbara, and 300 for you. This, as we have seen, are precisely the strategies that Barbara and you can rationally choose under common strong belief in rationality. \square

Example 9.6. Watching TV with Barbara.

Consider the graphical representation of this game in Figure 9.2. So, the three full decision problems in this game are $\Gamma^0(\emptyset)$, $\Gamma^0(h_1)$ and $\Gamma^0(h_2)$ as depicted in Table 9.7.



$$\Gamma^1(\emptyset): \text{ You active}$$

	(B, B)	(B, D)	(D, B)	(D, D)
$(fight, B)$	4, 1	4, 1	0, 0	0, 0
$(don't, B)$	6, 3	2, 2	6, 3	2, 2
$(don't, D)$	2, 2	3, 6	2, 2	3, 6

$$\Gamma^1(h_1): \text{ You and Barbara active}$$

	(B, B)	(B, D)	(D, B)	(D, D)
$(fight, B)$	4, 1	4, 1	0, 0	0, 0

$$\Gamma^1(h_2): \text{ You and Barbara active}$$

	(B, B)	(B, D)	(D, B)	(D, D)
$(don't, B)$	6, 3	2, 2	6, 3	2, 2
$(don't, D)$	2, 2	3, 6	2, 2	3, 6

Table 9.8: Reduced decision problems after step 1 in “Watching TV with Barbara”

Step 1. In the decision problem $\Gamma^0(\emptyset)$ at the beginning, only your strategy $(fight, D)$ is strictly dominated. We therefore eliminate $(fight, D)$ from $\Gamma^0(\emptyset)$ but also from the future decision problem $\Gamma^0(h_1)$. Namely, by doing so we do not remove all your strategies leading to h_1 , as $(fight, B)$ still remains. No other strategies can be eliminated in round 1. This leads to the reduced decision problems in Table 9.8.

Step 2. In decision problem $\Gamma^1(h_1)$, Barbara’s strategies (D, B) and (D, D) are strictly dominated. We therefore eliminate these two strategies from $\Gamma^1(h_1)$, but also from $\Gamma^1(\emptyset)$ and $\Gamma^1(h_2)$. No other strategies can be eliminated at step 2. This leads to the new decision problems in Table 9.9.

Step 3. In decision problem $\Gamma^2(\emptyset)$ at the beginning of the game, your strategy $(don't, D)$ is strictly dominated. We therefore eliminate this strategy from $\Gamma^2(\emptyset)$ but also from the future decision problem $\Gamma^2(h_2)$, as by doing so we do not remove all your strategies leading to h_2 . No other strategies can be eliminated in step 3. This leads to the new decision problems in Table 9.10.

Step 4. In decision problem $\Gamma^3(h_2)$, Barbara’s strategy (B, D) is strictly dominated by (B, B) . We thus remove Barbara’s strategy (B, D) from the decision problems $\Gamma^3(\emptyset)$, $\Gamma^3(h_1)$ and $\Gamma^3(h_2)$. No other strategies can



$\Gamma^2(\emptyset)$: You active

	(B, B)	(B, D)
$(fight, B)$	4, 1	4, 1
$(don't, B)$	6, 3	2, 2
$(don't, D)$	2, 2	3, 6

$\Gamma^2(h_1)$: You and Barbara active

	(B, B)	(B, D)
$(fight, B)$	4, 1	4, 1

$\Gamma^2(h_2)$: You and Barbara active

	(B, B)	(B, D)
$(don't, B)$	6, 3	2, 2
$(don't, D)$	2, 2	3, 6

Table 9.9: Reduced decision problems after step 2 in “Watching TV with Barbara”

$\Gamma^3(\emptyset)$: You active

	(B, B)	(B, D)
$(fight, B)$	4, 1	4, 1
$(don't, B)$	6, 3	2, 2

$\Gamma^3(h_1)$: You and Barbara active

	(B, B)	(B, D)
$(fight, B)$	4, 1	4, 1

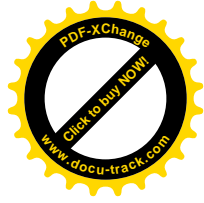
$\Gamma^3(h_2)$: You and Barbara active

	(B, B)	(B, D)
$(don't, B)$	6, 3	2, 2

Table 9.10: Reduced decision problems after step 3 in “Watching TV with Barbara”

be eliminated in step 4. This leads to the new decision problems in Table 9.11.

Step 5. In decision problem $\Gamma^4(\emptyset)$, your strategy $(fight, B)$ is strictly dominated by $(don't, B)$. We thus eliminate your strategy $(fight, B)$ from $\Gamma^4(\emptyset)$. However, we cannot eliminate this strategy from $\Gamma^4(h_1)$, since by



$$\Gamma^4(\emptyset): \text{ You active}$$

	(B, B)
$(fight, B)$	4, 1
$(don't, B)$	6, 3

$$\Gamma^4(h_1): \text{ You and Barbara active}$$

	(B, B)
$(fight, B)$	4, 1

$$\Gamma^4(h_2): \text{ You and Barbara active}$$

	(B, B)
$(don't, B)$	6, 3

Table 9.11: Reduced decision problems after step 4 in “Watching TV with Barbara”

$$\Gamma^5(\emptyset): \text{ You active}$$

	(B, B)
$(don't, B)$	6, 3

$$\Gamma^5(h_1): \text{ You and Barbara active}$$

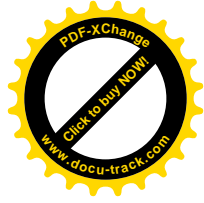
	(B, B)
$(fight, B)$	4, 1

$$\Gamma^5(h_2): \text{ You and Barbara active}$$

	(B, B)
$(don't, B)$	6, 3

Table 9.12: Final decision problems after step 5 in “Watching TV with Barbara”

doing so we would remove at h_1 all your strategies leading to h_1 . No further strategies can be eliminated at step 5. This leads to the final decision problems in Table 9.12, from which no strategies can be removed. So, the only strategies that survive the iterated conditional dominance procedure are the strategies $(don't, B)$ for you, and (B, B) for Barbara. We have seen that these are also the only strategies that can rationally be chosen under common strong belief in rationality. \square



9.4. Comparison with Backward Dominance Procedure

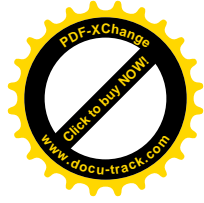
In this section we will compare in more detail the iterated conditional dominance procedure to the backward dominance procedure from Chapter 8. The backward dominance procedure is characterized by eliminating strategies backwards – that is, if a strategy s_i is strictly dominated in a decision problem at information set h where i is active, then we eliminate s_i at h , but also at all decision problems *before* h .

The iterated conditional dominance procedure works differently: If strategy s_i is strictly dominated in a decision problem at information set h where i is active, then we do not only eliminate s_i at h and all decision problems before h , but also at all decision problems *after* h – as long as we do not remove all strategy combinations leading to that information set. In other words, the iterated conditional dominance procedure eliminates *backwards and forward*.

Not only this, if s_i is strictly dominated at h , then it also eliminates s_i at information sets that do not come before, nor after, h . Consider, for instance, the example “Watching TV with Barbara”. In step 2 of the iterated conditional dominance procedure, Barbara’s strategies (D, B) and (D, D) are strictly dominated at h_1 , and we thus eliminate these strategies also at h_2 , which is an information set that does not come before, nor after, h_1 . We say that h_1 and h_2 are “parallel” information sets. We can thus say that the iterated conditional dominance procedure works by eliminating strategies backwards, forward, and in parallel.

At first sight, this could suggest that in the iterated conditional dominance procedure we always eliminate more strategies at every step than in the backward dominance procedure, and that therefore the iterated conditional dominance procedure is always more restrictive. This is not true, however! Note that in the iterated conditional dominance procedure we always have the “unless” at every elimination step. Namely, from the decision problem $\Gamma^k(h)$ at h we eliminate the strategies as specified in the algorithm *unless* this would remove all strategy combinations leading to h . In the latter case, we do not remove any strategies from $\Gamma^k(h)$. At the same time, it is possible that in the latter case we would still eliminate strategies at h in the backward dominance procedure. So, it is not true that the iterated conditional dominance procedure always eliminates more than the backward dominance procedure.

In fact, there is no logical relationship, in terms of strategy choices, between the outputs of the iterated conditional dominance procedure and the backward dominance procedure – sometimes the former is more restrictive, sometimes the latter, and it may also happen that both procedures yield completely opposed sets of strategies for a given player.



Consider, for instance, the example “Painting Chris’ house” with the graphical representation in Figure 9.1. We have seen that the iterated conditional dominance procedure uniquely selects the strategy 300 for you here, whereas the backward dominance procedure uniquely selects the strategy 200 for you. Hence, in this example the two procedures yield completely opposed results.

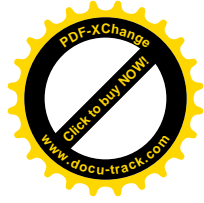
Now, suppose that in this example Barbara would receive 500 instead of 350 by accepting the colleague’s offer. Then, it may be checked that the iterated conditional dominance procedure would select the strategies 200, 300 and 400 for you, whereas the backward dominance procedure would still uniquely select your strategy 200. So, in this modified example, the backward dominance procedure would be more restrictive in terms of strategies for you.

Next, consider the example “Watching TV with Barbara”. As we have seen, the iterated conditional dominance procedure uniquely filters the strategy $(don't, B)$ for you. The reader may verify that the backward dominance procedure only eliminates the strategy $(fight, D)$ for you at $\Gamma^0(\emptyset)$, after which the procedure terminates. Hence, the backward dominance procedure selects the strategies $(fight, B)$, $(don't, B)$ and $(don't, D)$ for you. That is, in this example the iterated conditional dominance procedure is more restrictive in terms of possible strategy choices for you.

Overall, we may thus conclude that, in terms of strategy choices, there is no general logical relationship between the iterated conditional dominance procedure and the backward dominance procedure.

What can we say about dynamic games with *perfect information* – that is, games in which players choose one at the time, and always observe the choices made by their opponents in the past? We have seen in the previous chapter that in such games, the backward dominance procedure is equivalent to the *backward induction procedure*. As a consequence, in a game with perfect information the strategies that can rationally be chosen under common belief in future rationality are precisely the backward induction strategies.

Is this also true for common strong belief in rationality and the associated iterated conditional dominance procedure? The answer is “no”! To show this, we will now give an example of a dynamic game with perfect information, in which the strategies that can rationally be chosen under common strong belief in rationality are different from the backward induction strategies. Even stronger, for one of the players there is a



unique strategy that can rationally be chosen under common strong belief in rationality, and this strategy is different from the unique backward induction strategy in the game.

Example 9.7. The heat of the fight.

It is Wednesday evening, and you and Barbara face the weekly problem of deciding which program to watch on TV: *Blackadder* or *Dallas*. Remember that you prefer *Blackadder*, whereas Barbara prefers *Dallas*. Suppose, as before, that watching *Blackadder* would give you a utility of 6 and Barbara a utility of 3, whereas for watching *Dallas* it is the other way around. At the beginning of the evening, Barbara has the choice to either be *nice* to you and let you choose your favorite program, or to *argue* with you about the program. If she starts to *argue* then you can either be *nice* to her and let her choose her favorite program, in order to avoid any escalation of the conflict, or you can decide to *shout* at her as a response. If you decide to *shout* at her then Barbara has the option to be *nice* to you and let you choose your favorite program, so as to prevent the situation from getting out of hand, or she can start throwing *dishes* on the floor as a sign of her anger. If she throws *dishes* on the floor then, because the situation has gone completely out of hand, you can only choose between two extremes – you either apologize to her and say that you are *sorry* about this whole fight, and let her choose her favorite program in addition, or you walk *out* the door and watch *Blackadder* at Chris' house.

Assume that the utility for both you and Barbara would decrease by 5 every time the conflict escalates. However, if you apologize to Barbara at the end then this would increase Barbara's utility by 15. Moreover, if you decide to walk out the door and visit Chris, then this would increase your utility by 15 since Chris will serve some nice beer and potatoe chips to make you feel better. This situation can be represented graphically by Figure 9.3. Here, the first utility corresponds to Barbara (player 1) and the second utility to you (player 2). Note that if you walk out the door at the end, then Barbara will still watch her favorite program *Dallas* but the conflict would have escalated three times. So, her utility in that case is $6 - 15 = -9$.

Let us first analyze which strategy, or strategies, you can rationally choose under common belief in future rationality. Since this is a dynamic game with perfect information, we can find these strategies by performing the backward induction procedure. So we start at the end of the game, where you must choose between saying *sorry* and walking *out* the door. Clearly, the optimal choice for you there is to walk *out*.

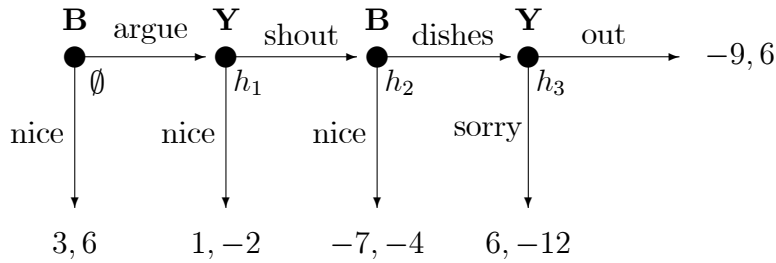


Figure 9.3: The heat of the fight

We then move to the penultimate information set, where Barbara must choose between being *nice* to you and throwing *dishes* on the floor. The optimal choice here for Barbara, given that we have selected the unique backward induction choice *out* for you at the final information set, is to be *nice* to you.

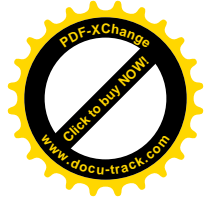
Subsequently, we analyze the information set where you must choose between being *nice* to Barbara or to *shout* at her. Given that we have selected the backward induction choice *nice* for Barbara at the information set that follows, the optimal choice for you here is to be *nice* to Barbara.

Finally, we investigate the beginning of the game. As we have selected the backward induction choice *nice* for you at the information set that follows, the backward induction choice for Barbara at the beginning of the evening is to be *nice* to you, and let you choose your favorite program.

So, we see that the unique backward induction strategy for you in this game is to be *nice* to Barbara in case she starts arguing about the program. Hence, the only strategy that you can rationally choose under common belief in future rationality is to be *nice* to Barbara.

We will now explore which strategy, or strategies, you can rationally choose under common strong belief in rationality. To answer this question we use the iterated conditional dominance procedure. Let us denote the four information sets in this game by \emptyset (the beginning), h_1 , h_2 and h_3 . The full decision problems at these four information sets are given in Table 9.13. Here, the strategies for Barbara have been put in the rows, and the strategies for you in the columns.

Step 1. Note that in $\Gamma^0(\emptyset)$, Barbara's strategy (*argue*, *nice*) is strictly dominated by her strategy *nice*. So, we eliminate her strategy (*argue*, *nice*) in the decision problems at \emptyset , h_1 and h_2 . Moreover, in $\Gamma^0(h_1)$ your strategy (*shout*, *sorry*) is strictly dominated by your strategy *nice*, and in $\Gamma^0(h_3)$



$\Gamma^0(\emptyset)$: Barbara active

	<i>nice</i>	<i>(shout, sorry)</i>	<i>(shout, out)</i>
<i>nice</i>	3, 6	3, 6	3, 6
<i>(argue, nice)</i>	1, -2	-7, -4	-7, -4
<i>(argue, dishes)</i>	1, -2	6, -12	-9, 6

$\Gamma^0(h_1)$: You active

	<i>nice</i>	<i>(shout, sorry)</i>	<i>(shout, out)</i>
<i>(argue, nice)</i>	1, -2	-7, -4	-7, -4
<i>(argue, dishes)</i>	1, -2	6, -12	-9, 6

$\Gamma^0(h_2)$: Barbara active

	<i>(shout, sorry)</i>	<i>(shout, out)</i>
<i>(argue, nice)</i>	-7, -4	-7, -4
<i>(argue, dishes)</i>	6, -12	-9, 6

$\Gamma^0(h_3)$: You active

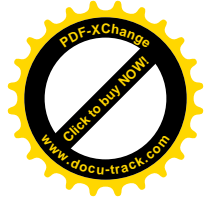
	<i>(shout, sorry)</i>	<i>(shout, out)</i>
<i>(argue, dishes)</i>	6, -12	-9, 6

Table 9.13: Full decision problems in “The heat of the fight”

the same strategy *(shout, sorry)* for you is strictly dominated by your strategy *(shout, out)*. Hence, we eliminate your strategy *(shout, sorry)* from the decision problems at \emptyset, h_1, h_2 and h_3 . This leads to the reduced decision problems in Table 9.14.

Step 2. In the reduced decision problem $\Gamma^1(\emptyset)$ at the beginning, Barbara’s strategy *(argue, dishes)* is strictly dominated by her strategy *nice*. So, we eliminate her strategy *(argue, dishes)* from $\Gamma^1(\emptyset)$ but *not* from the decision problems at h_1, h_2 and h_3 , since otherwise we would delete all of Barbara’s strategies from the reduced decision problems at h_1, h_2 and h_3 . Moreover, in $\Gamma^1(h_1)$ your strategy *nice* is strictly dominated by your strategy *(shout, out)*, and hence we eliminate your strategy *nice* from $\Gamma^1(\emptyset)$ and $\Gamma^1(h_1)$. This leads to the reduced decision problems in Table 9.15, from which no further strategies can be eliminated.

So, the only strategies that survive the iterated conditional dominance procedure are the strategy *nice* for Barbara, and the strategy *(shout, out)* for you. As a consequence, *(shout, out)* is the only strategy you can rationally choose under common strong belief in rationality. This is different



$$\Gamma^1(\emptyset): \text{Barbara active}$$

	<i>nice</i>	<i>(shout, out)</i>
<i>nice</i>	3, 6	3, 6
<i>(argue, dishes)</i>	1, -2	-9, 6

$$\Gamma^1(h_1): \text{You active}$$

	<i>nice</i>	<i>(shout, out)</i>
<i>(argue, dishes)</i>	1, -2	-9, 6

$$\Gamma^1(h_2): \text{Barbara active}$$

	<i>(shout, out)</i>
<i>(argue, dishes)</i>	-9, 6

$$\Gamma^1(h_3): \text{You active}$$

	<i>(shout, out)</i>
<i>(argue, dishes)</i>	-9, 6

Table 9.14: Reduced decision problems after step 1 of iterated conditional dominance procedure in “The heat of the fight”

$$\Gamma^2(\emptyset): \text{Barbara active}$$

	<i>(shout, out)</i>
<i>nice</i>	3, 6

$$\Gamma^2(h_1): \text{You active}$$

	<i>(shout, out)</i>
<i>(argue, dishes)</i>	-9, 6

$$\Gamma^2(h_2): \text{Barbara active}$$

	<i>(shout, out)</i>
<i>(argue, dishes)</i>	-9, 6

$$\Gamma^2(h_3): \text{You active}$$

	<i>(shout, out)</i>
<i>(argue, dishes)</i>	-9, 6

Table 9.15: Final decision problems in the iterated conditional dominance procedure in “The heat of the fight”



from the unique strategy you can rationally choose under common belief in future rationality, which was the strategy *nice* as we have seen.

There is also an easy intuitive argument for why common belief in future rationality and common strong belief in rationality lead to different strategy choices for you in this example. Suppose, namely, that at h_1 you observe that Barbara has started to *argue* with you about the TV program. Under common belief in future rationality, you believe that Barbara will choose rationally at h_2 , and you believe that Barbara believes that you will choose rationally at h_3 . Your only optimal choice at h_3 is to walk *out*, so you believe at h_1 that Barbara believes at h_2 that you will choose *out* at h_3 . If Barbara indeed believes so, her optimal choice at h_2 would be to be *nice* to you. Since you believe at h_1 that Barbara chooses rationally at h_2 , you believe at h_1 that Barbara will indeed be *nice* to you at h_2 , and hence you will be *nice* to her at h_1 . So, under common belief in future rationality, your optimal strategy is to be *nice* to Barbara in case she starts to *argue*.

Under common strong belief in rationality your reasoning will be crucially different at h_1 . Suppose that at h_1 you observe that Barbara has started to *argue* about the program, and assume you strongly believe in Barbara's rationality at h_1 . Then you ask whether her choice to *argue* could be part of an optimal strategy for Barbara. This is indeed possible, but only if she would subsequently throw *dishes* on the floor, hoping for you to apologize afterwards. This is the only way for her to get a utility higher than 3 – the utility she gets by being *nice* to you at the beginning. Hence, if you strongly believe in Barbara's rationality, then you must believe at h_1 that Barbara is implementing the strategy (*argue, dishes*), and therefore your only optimal strategy choice at h_1 is to choose (*shout, out*). So, under common strong belief in rationality, the optimal thing for you to do is to *shout* if she starts arguing with you, and walk *out* the door in case she throws dishes on the floor. \square

The example above thus shows that even in dynamic games with perfect information, the concepts of common strong belief in rationality and common belief in future rationality – or equivalently the iterated conditional dominance procedure and the backward induction procedure – may yield completely opposed strategy choices for you. Hence, in terms of strategy choices there is no general logical relationship between the two concepts, even if we restrict to dynamic games with perfect information.

Things are different, however, if we focus on the outcomes – that is, the terminal histories – you deem possible before the game starts. In



each of the examples we discussed above, every outcome you deem possible before the game starts under common strong belief in rationality, is also initially deemed possible under common belief in future rationality! In the example “Painting Chris’ house”, for instance, there is only one outcome you initially deem possible under common strong belief in rationality, and under common belief in future rationality, namely that Barbara will accept the colleague’s offer at the beginning. The same holds if in this example Barbara would receive 500 instead of 350 by accepting the colleague’s offer. Note, however, that both concepts differ in the restrictions on your possible choices if you surprisingly observe that Barbara rejects the colleague’s offer – something you initially believe not to happen. In the example “Watching TV with Barbara” there is only one outcome you initially deem possible under common strong belief in rationality, namely that you will not start a fight with Barbara, and that you will both watch your favorite program *Blackadder* together. This is also one of the outcomes you initially deem possible under common belief in future rationality. In the example “The heat of the fight”, the only outcome you initially deem possible under common strong belief in rationality, and under common belief in future rationality, is that Barbara will be *nice* to you at the beginning of the game. We have seen, however, that both concepts differ in the restrictions on your behavior if you surprisingly observe that Barbara starts to *argue* with you – something you initially believe not to happen.

In fact, we can show that this is always true. That is, in every dynamic game, any outcome you initially deem possible under common strong belief in rationality, is also initially deemed possible under common belief in future rationality. Before we state this result formally, we first define precisely what we mean by “outcomes that are initially deemed possible under common strong belief in rationality, and under common belief in future rationality”.

DEFINITION 9.4.1. (*Outcomes that are initially deemed possible*)

You **initially** deem an outcome z **possible** under common strong belief in rationality (common belief in future rationality) if there is a strategy combination leading to z where every strategy can rationally be chosen under common strong belief in rationality (common belief in future rationality).

Here, by an outcome z we formally mean a terminal history z in the dynamic game – that is, a situation where the game ends. Note that under common strong belief in rationality you believe, before the start of the game, that every opponent will choose a strategy that can rationally



be chosen under common strong belief in rationality, and the same applies to common belief in future rationality. So, the definition above matches exactly our intuition. We can now formally state the result we announced above.

THEOREM 9.4.2. (*Outcomes under common strong belief in rationality and common belief in future rationality*)

Every outcome you initially deem possible under common strong belief in rationality, is also initially deemed possible under common belief in future rationality.

The proof for this result can be found in the proofs section at the end of this chapter. Note that the converse of this theorem is not true: In the example “Watching TV with Barbara”, there is only one outcome you initially deem possible under common strong belief in rationality, namely that you will not start a fight with Barbara, and that you will both watch *Blackadder*. However, under common belief in future rationality you also deem possible other outcomes initially – for instance that you will start a fight with Barbara, that you will choose *Blackadder*, and that Barbara chooses *Dallas*.

The theorem above has important consequences for games with perfect information. We have seen in Chapter 8 that in games with perfect information, the strategies that can rationally be chosen under common belief in future rationality are exactly the backward induction strategies. Now, call an outcome in such a game a *backward induction outcome* if there is a combination of backward induction strategies that leads to this outcome. So, in a game with perfect information, the outcomes you initially deem possible under common belief in future rationality are precisely the backward induction outcomes. By the theorem above, we may thus conclude that in games with perfect information, every outcome you initially deem possible under common strong belief in rationality must be a backward induction outcome.

COROLLARY 9.4.3. (*Common strong belief in rationality leads to backward induction outcomes*)

In a game with perfect information, every outcome you initially deem possible under common strong belief in rationality must be a backward induction outcome.

However, under common strong belief in rationality you may no longer reason in accordance with backward induction if you surprisingly observe that your opponent has not chosen in accordance with backward induction before – something you initially believe not to happen. Consider,



for instance, the example “The heat of the fight”, where you initially believe that Barbara will make the backward induction choice to be *nice* to you at the beginning. However, if you observe at h_1 that Barbara has not chosen the backward induction strategy *nice*, then under common strong belief in rationality you will believe that she would subsequently make the non-backward induction choice to throw *dishes* on the floor. Consequently, you would choose (*shout, out*) rather than your backward induction strategy *nice*.

9.5. Order Dependence

In this section we will discuss the issue of changing the order and speed of elimination in the iterated conditional dominance procedure. A very important property of the backward dominance procedure, as we have seen in Chapter 8, is that its outcome does not depend on the order and speed of elimination. Even if we do not eliminate, at a given round and at a given information set, all strategies we can, then this will not affect the final outcome of the procedure – at the end, the set of strategies that survive will be exactly the same as when we would always eliminate everything we can at every round and every information set.

This no longer holds for the iterated conditional dominance procedure, however! For this procedure to give the “correct” result, it is absolutely crucial that at every round and every information set, we always eliminate *all* strategies that we can. Otherwise, the set of surviving strategies at the end could change. As an illustration, consider the example “Painting Chris’ house”.

Example 9.8. Painting Chris’ house.

Consider the graphical representation in Figure 9.1, and the full decision problems $\Gamma^0(\emptyset)$ and $\Gamma^0(h_1)$ in Table 9.4. We have seen that for Barbara, the strategies $(r, 200)$, $(r, 300)$ and $(r, 500)$ are all strictly dominated by *accept* in the decision problem $\Gamma^0(\emptyset)$. Therefore, according to the original iterated conditional dominance procedure, we must eliminate each of these strategies from $\Gamma^0(\emptyset)$ and $\Gamma^0(h_1)$. Suppose now that from the decision problem $\Gamma^0(h_1)$ at h_1 we would *only* eliminate the strategies $(r, 300)$ and $(r, 500)$ for Barbara, but not the strategy $(r, 200)$, in step 1. Assume that we would still remove your strategy 500 from $\Gamma^0(\emptyset)$ and $\Gamma^0(h_1)$, as it is strictly dominated in $\Gamma^0(h_1)$. This would lead to the reduced decision problems $\hat{\Gamma}^1(\emptyset)$ and $\hat{\Gamma}^1(h_1)$ in Table 9.16.

In the decision problem $\hat{\Gamma}^1(\emptyset)$, Barbara’s strategy $(r, 400)$ is strictly dominated, and hence we can eliminate it at $\hat{\Gamma}^1(\emptyset)$ and $\hat{\Gamma}^1(h_1)$. Note,



$$\hat{\Gamma}^1(\emptyset): \text{Barbara active}$$

	200	300	400
$(r, 400)$	0, 200	0, 300	200, 200
<i>accept</i>	350, 500	350, 500	350, 500

$$\hat{\Gamma}^1(h_1): \text{Barbara and you active}$$

	200	300	400
$(r, 200)$	100, 100	200, 0	200, 0
$(r, 400)$	0, 200	0, 300	200, 200

Table 9.16: Changing order of elimination in “Painting Chris’ house”:
Step 1

$$\hat{\Gamma}^2(\emptyset): \text{Barbara active}$$

	200	300
<i>accept</i>	350, 500	350, 500

$$\hat{\Gamma}^2(h_1): \text{Barbara and you active}$$

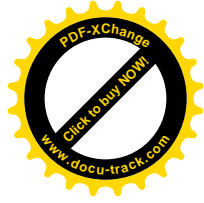
	200	300
$(r, 200)$	100, 100	200, 0

Table 9.17: Changing order of elimination in “Painting Chris’ house”:
Step 2

namely, that by doing so we will not eliminate all of Barbara’s strategies leading to h_1 , as $\hat{\Gamma}^1(h_1)$ still contains another strategy $(r, 200)$ for Barbara. Moreover, at $\hat{\Gamma}^1(h_1)$, your strategy 400 is strictly dominated by the randomized strategy $(0.5) \cdot 200 + (0.5) \cdot 300$, and therefore we can eliminate your strategy 400 in $\hat{\Gamma}^1(\emptyset)$ and $\hat{\Gamma}^1(h_1)$. This leads to the new decision problems $\hat{\Gamma}^2(\emptyset)$ and $\hat{\Gamma}^2(h_1)$ in Table 9.17.

In the decision problem $\hat{\Gamma}^2(h_1)$, your strategy 300 is strictly dominated, and hence we can eliminate it from $\hat{\Gamma}^2(\emptyset)$ and $\hat{\Gamma}^2(h_1)$. This would lead to the final decision problems in Table 9.18.

So, by changing the order of elimination in this way, the unique strategy for you that would survive is 200. This, however, is different from the unique strategy for you that survives the *original* iterated conditional dominance procedure – namely 300. Hence, changing the order and speed



$$\hat{\Gamma}^3(\emptyset): \text{Barbara active}$$

	200
<i>accept</i>	350, 500

$$\hat{\Gamma}^3(h_1): \text{Barbara and you active}$$

	200
<i>(r, 200)</i>	100, 100

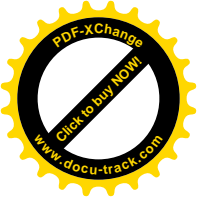
Table 9.18: Changing order of elimination in “Painting Chris’ house”:
Step 3

of elimination may drastically change the outcome of the iterated conditional dominance procedure! \square

This example thus shows that, if we use the iterated conditional dominance procedure, there is no other way than to eliminate, at every round and every information set, all strategies we can. If we do not, we run the danger of ending up with a different set of strategies. In particular, we cannot look for a “most convenient order of elimination” – as we often did for the backward dominance procedure – when working with the iterated conditional dominance procedure. For instance, it is no longer allowed to use the “backwards order of elimination” which turned out to be very useful for the backward dominance procedure in many games. This makes the iterated conditional dominance procedure somewhat harder to work with than the backward dominance procedure, especially when the dynamic game is large.

9.6. Rationality Orderings

In a sense, the concept of common strong belief in rationality groups the strategies of every player into classes, ranging from “most rational” to “least rational”, with typically some classes in between. First, we have the strategies that cannot be optimal under any conditional belief vector – the “least rational strategies” or 0-rational strategies. Then, we have the strategies that are optimal for some conditional belief vector, but that cannot be rationally chosen under 1-fold strong belief in rationality. We call these strategies the 1-rational strategies, as they are “more rational” than the 0-rational strategies – they can rationally be chosen under *some* conditional belief vector – but less rational than those which can rationally be chosen under 1-fold strong belief in rationality. We then have the 2-rational strategies, which are the ones that can rationally be



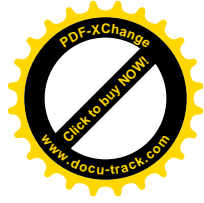
chosen under 1-fold strong belief in rationality, but not under expressing up to 2-fold strong belief in rationality, and so on.

This process must stop some time, as for every game there is a number K such that common strong belief in rationality is equivalent to expressing up to K -fold strong belief in rationality. So, in this way we obtain for every player i an ordering $(D_i^0, D_i^1, \dots, D_i^K)$ of his strategies into subclasses, where D_i^0 is the set of 0-rational strategies, D_i^1 is the set of 1-rational strategies, and so on, until we reach D_i^K , which is the class of K -rational, or “most rational” strategies. For every $k \in \{1, \dots, K - 1\}$, the class D_i^k is the set of k -rational strategies, and it contains those strategies that can rationally be chosen by a player who expresses up to $(k - 1)$ -fold strong belief in rationality, but not by a player who expresses up to k -fold strong belief in rationality. The class D_i^K – containing the “most rational” strategies – consists of those strategies that can rationally be chosen under common strong belief in rationality. Finally, the class D_i^0 – the “least rational” strategies – contains those strategies that are not optimal for any conditional belief vector. Such an ordering of strategies into subclasses is called a *rationality ordering*.

The rationality ordering $(D_i^0, D_i^1, \dots, D_i^K)$ induced by common strong belief in rationality has a clear intuitive meaning: For every $k \in \{1, \dots, K - 1\}$, the k -rational strategies are “more rational” than the $(k - 1)$ -rational strategies, but “less rational” than the $(k + 1)$ -rational strategies. So, it orders player i 's strategies from “most rational” to “least rational”, with usually some levels in between. Of course, the terms “more rational”, “less rational”, “most rational” and “least rational” are all subjective, as it reflects the particular viewpoint taken by the concept of common strong belief in rationality.

As an illustration, let us go back to the example “Watching TV with Barbara”. As we know by Theorem 9.3.3, there is an easy way to find those strategies that player i can rationally choose by expressing up to k -fold strong belief in rationality. These, namely, are precisely the strategies in $\Gamma^{k+1}(\emptyset)$ – the reduced decision problem at the beginning of the game after the first $k + 1$ rounds of the iterated conditional dominance procedure. In Example 9.6 we have computed these decision problems $\Gamma^k(\emptyset)$, and for our convenience we have reproduced the result in Table 9.19.

With the help of these decision problems $\Gamma^k(\emptyset)$ we can now easily derive the rationality ordering induced on your strategies. The set D_1^0 – containing your “least rational” or 0-rational strategies – are those strategies that are not optimal for any conditional belief vector. These are precisely your strategies which are in $\Gamma^0(\emptyset)$ but not in $\Gamma^1(\emptyset)$, which in this case is only your strategy $(fight, D)$. So, $D_1^0 = \{(fight, D)\}$.



	Your strategies	Barbara's strategies
$\Gamma^0(\emptyset)$	$(fight, B), (fight, D), (don't, D), (don't, B)$	$(B, B), (B, D), (D, B), (D, D)$
$\Gamma^1(\emptyset)$	$(fight, B), (don't, D), (don't, B)$	$(B, B), (B, D), (D, B), (D, D)$
$\Gamma^2(\emptyset)$	$(fight, B), (don't, D), (don't, B)$	$(B, B), (B, D)$
$\Gamma^3(\emptyset)$	$(fight, B), (don't, B)$	$(B, B), (B, D)$
$\Gamma^4(\emptyset)$	$(fight, B), (don't, B)$	(B, B)
$\Gamma^5(\emptyset)$	$(don't, B)$	(B, B)

Table 9.19: The reduced decision problems $\Gamma^k(\emptyset)$ in the example “Watching TV with Barbara”

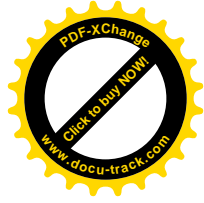
The set D_1^1 of 1-rational strategies contains those strategies for you that can rationally be chosen under some conditional belief vector, but not if you express 1-fold strong belief in rationality. These are the strategies for you that are in $\Gamma^1(\emptyset)$ but not in $\Gamma^2(\emptyset)$, which are none. So, the set D_1^1 of 1-rational strategies is empty.

The set D_1^2 of 2-rational strategies contains those strategies for you that can rationally be chosen under expressing 1-fold strong belief in rationality, but not under expressing up to 2-fold strong belief in rationality. These are precisely the strategies for you that are in $\Gamma^2(\emptyset)$ but not in $\Gamma^3(\emptyset)$, that is, your strategy $(don't, D)$. So, $D_1^2 = \{(don't, D)\}$. By continuing in this fashion, you will find that D_1^3 is empty and that $D_1^4 = \{(fight, B)\}$.

Let us finally focus on D_1^5 – the set of “most rational” or 5-rational strategies for you. This set contains those strategies that you can rationally choose under common strong belief in rationality. These are precisely the strategies for you in $\Gamma^5(\emptyset)$ – that is, $(don't, B)$ – as the iterated conditional dominance procedure stops after step 5. Hence, $D_1^5 = \{(don't, B)\}$.

So, the rationality ordering on your strategies induced by common strong belief in rationality is

$$(\{(fight, D)\}, \emptyset, \{(don't, D)\}, \emptyset, \{(fight, B)\}, \{(don't, B)\}).$$



In a similar way, the reader may verify that the rationality ordering on Barbara's strategies is given by

$$(\emptyset, \{(D, B), (D, D)\}, \emptyset, \{(B, D)\}, \emptyset, \{(B, B)\}).$$

That is, the set of 0-rational strategies for Barbara is empty, the set of 1-rational strategies is $\{(D, B), (D, D)\}$, the set of 2-rational strategies is empty, the set of 3-rational strategies is $\{(B, D)\}$, the set of 4-rational strategies is empty, and the set of 5-rational, or "most rational", strategies is $\{(B, B)\}$.

In general, a rationality ordering for player i in an arbitrary dynamic game is an ordering of his strategies into subclasses, ranging from the "least rational" to the "most rational strategies". Here is a formal definition.

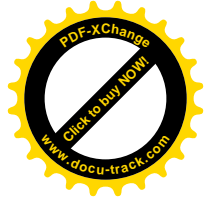
DEFINITION 9.6.1. (*Rationality ordering*)

A **rationality ordering** for player i in a dynamic game is a sequence $R_i = (D_i^0, D_i^1, \dots, D_i^K)$ where D_i^0, \dots, D_i^K are disjoint sets of strategies for player i whose union is equal to the full set of strategies S_i . Here, some of these sets may be empty, but not D_i^K . For every k , all strategies in D_i^k are called **k -rational** strategies. The interpretation is that D_i^0 contains the "least rational" strategies, D_i^K the "most rational strategies", and all strategies in D_i^{k+1} are "more rational" than all strategies in D_i^k .

Here, by "disjoint" we mean that the sets D_i^0, \dots, D_i^K have no overlap. That is, every strategy for player i is in exactly one of the sets from D_i^0, \dots, D_i^K .

Above we have seen how the concept of common strong belief in rationality naturally induces a rationality ordering for every player. This is just one example of a rationality ordering, however, as a rationality ordering may in fact be *any* ordering of strategies according to the definition above. But not all rationality orderings make intuitive sense! If one takes a rationality ordering $R_i = (D_i^0, D_i^1, \dots, D_i^K)$, then there must be a good reason for why D_i^0 contains the "least rational" strategies, why D_i^K contains the "most rational strategies", and why strategies in D_i^{k+1} are deemed "more rational" than strategies in D_i^k .

Throughout this book, we have taken the perspective that "more rational" strategies are the ones that are supported by "more rational" beliefs. Now, in order to classify the beliefs that player i has about his opponents' strategy combinations, we must look at the opponents' rationality orderings of strategies. Suppose that for every opponent j we have a rationality ordering $R_j = (D_j^0, D_j^1, \dots, D_j^K)$ of his strategies. Then, a "most rational" belief for player i assigns at every information



set $h \in H_i$ only positive probability to “most rational” strategies by the opponents, if possible. That is, if at information set $h \in H_i$ there is a combination of opponents’ strategies leading to h consisting only of “most rational” strategies, then a “most rational” belief should at h only assign positive probability to opponents’ strategy combinations consisting of “most rational” strategies. Moreover, if there is no combination of opponents’ strategies leading to h consisting only of “most rational” strategies, but there is a combination of opponents’ strategies leading to h consisting of at-least- $(K - 1)$ -rational strategies, then a “most rational” belief for player i should at h only assign positive probability to opponents’ strategy combinations consisting of at-least- $(K - 1)$ -rational strategies. Here, by “at-least- $(K - 1)$ -rational” we mean $(K - 1)$ -rational or K -rational. And if at h there is no combination of opponents’ strategies leading to h consisting of at-least- $(K - 1)$ -rational strategies, but there is a combination of opponents’ strategies leading to h consisting of at-least- $(K - 2)$ -rational strategies, then a “most rational” belief for player i should at h only assign positive probability to opponents’ strategy combinations consisting of at-least- $(K - 2)$ -rational strategies, and so on. That is, a “most rational” belief for player i looks at every information set h for the highest degree k such that there is an opponents’ combination of at-least- k -rational strategies leading to h , and assigns at h only positive probability to at-least- k -rational strategies for the opponents. In other words, at every information set $h \in H_i$ player i attributes the highest possible rationality level to his opponents, according to their respective rationality orderings. We say that player i *strongly believes in the opponents’ rationality orderings*.

DEFINITION 9.6.2. (*Strong belief in the opponents’ rationality orderings*)

Consider a player i in a dynamic game. For every opponent j , let $R_j = (D_j^0, D_j^1, \dots, D_j^K)$ be a rationality ordering, with the same K for every opponent. At information set $h \in H_i$, let k be the highest degree such that there is an opponents’ combination of at-least- k -rational strategies leading to h . Then, a conditional belief vector for player i **strongly believes in the opponents’ rationality orderings** at h if it assigns at h only positive probability to opponents’ combinations of at-least- k -rational strategies.

A conditional belief vector for player i *strongly believes in the opponents’ rationality orderings* if it does so at every information set $h \in H_i$.



Here, by “at-least- k -rational” we mean a strategy that is k -rational or higher. So, our requirement will be that a “most rational” belief for player i should strongly believe in the opponents’ rationality orderings.

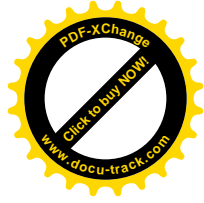
But what about “less rational” beliefs, which would support “less rational” strategies? We can weaken the requirement of strong belief in the opponents’ rationality orderings in the following way: Choose some level $m < K$, and require that player i only strongly believes in the opponents’ rationality orderings *up to level m* . More precisely, at every information set $h \in H_i$ we first look whether there is an opponents’ combination of at-least- m -rational strategies that leads to h . If so, then we require player i at h to only assign positive probability to at-least- m -rational strategies for the opponents, but *not necessarily to K -rational strategies*. So, even if there is an opponents’ combination of K -rational strategies leading to h , we only require player i to restrict at h to at-least- m -rational strategies for the opponents. If there is no opponents’ combination of at-least- m -rational strategies leading to h , we look whether there is an opponents’ combination of at-least- $(m - 1)$ -rational strategies leading to h . If so, then we require player i at h to only assign positive probability to at-least- $(m - 1)$ -rational strategies for the opponents. And so on. This leads to the following formal definition.

DEFINITION 9.6.3. (*Strong belief in the opponents’ rationality orderings up to level m*)

Consider a player i in a dynamic game. For every opponent j , let $R_j = (D_j^0, D_j^1, \dots, D_j^K)$ be a rationality ordering, with the same K for every opponent. Fix a level $m \leq K$. At information set $h \in H_i$, let k be the highest degree **less than or equal to m** such that there is an opponents’ combination of at-least- k -rational strategies leading to h . Then, a conditional belief vector for player i **strongly believes in the opponents’ rationality orderings up to level m** at h if it assigns at h only positive probability to opponents’ combinations of at-least- k -rational strategies.

A conditional belief vector for player i strongly believes in the opponents’ rationality orderings up to level m if it does so at every information set $h \in H_i$.

So, one could say that under this weaker requirement, player i views all opponents’ strategies which are m -rational or higher as “equally rational”. For instance, if $m = K - 2$, then player i does not make a distinction between opponents’ strategies that are $(K - 2)$ -rational, or $(K - 1)$ -rational, or K -rational – they are all viewed as being “equally rational” by player i . Note that strong belief in the opponents’ rationality



orderings, as formulated in the previous definition, is the same as strong belief in the opponents' rationality orderings up to level K – the highest possible level. At the other extreme, by choosing $m = 0$, we impose no conditions at all on the conditional belief vectors for player i .

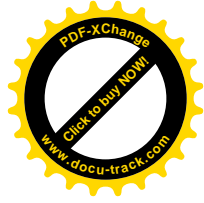
With the above definition at hand, we can now naturally classify the players' beliefs from “most rational” to “least rational”. For every player, fix a rationality ordering on strategies with maximum level K . Then, the “most rational” belief vectors for player i would be those that strongly believe in the opponents' rationality orderings up to level K . The “least rational” belief vectors would be those that strongly believe in the opponents' rationality orderings up to level 0, but not up to level 1. And for every level k in between, the k -rational belief vectors are those that strongly believe in the opponents' rationality orderings up to level k , but not up to level $k + 1$.

This, in turn, induces a natural classification of player i 's strategies from “most rational” to “least rational”! Namely, player i 's “most rational” strategies would intuitively be those that are optimal under a “most rational” belief vector – that is, a belief vector that strongly believes in the opponents' rationality orderings up to level K . Moreover, for every k between 1 and $K - 1$, the k -rational strategies for player i are intuitively those that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level $k - 1$, but not up to level k .

Summarizing, we see that if we start with a rationality ordering $R_i = (D_i^0, D_i^1, \dots, D_i^K)$ for every player i , with the same maximum level K for all players, then this induces for every player i a classification of his conditional belief vectors, from 0-rational (“least rational”) to K -rational (“most rational”), which in turns induces for every player i a classification of his strategies from 0-rational to K -rational. But then, intuitively, the latter classification of strategies for player i from 0-rational to K -rational must be the same as the classification in the rationality ordering R_i we started with! Combinations of rationality orderings with this property are called *self-confirming*, as the *initial* classifications of strategies given by the original rationality orderings are *confirmed* by the *induced* classifications of strategies as described above. So, in a sense, the self-confirming combinations of rationality orderings are those where the classifications of strategies from “most rational” to “least rational” are well-founded.

DEFINITION 9.6.4. (*Self-confirming combinations of rationality orderings*)

Consider for every player i a rationality ordering $R_i = (D_i^0, D_i^1, \dots, D_i^K)$



on his strategies, with the same maximum level K for every player. Then, this combination of rationality orderings is **self-confirming** if for every player i ,

the set D_i^K contains exactly those strategies that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level K ,

for every k between 1 and $K - 1$, the set D_i^k contains exactly those strategies that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level $k - 1$, but not under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level k , and

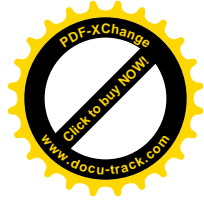
the set D_i^0 contains those strategies that are not optimal under any conditional belief vector.

The reader may verify that in the example “Watching TV with Barbara”, the rationality orderings induced by common strong belief in rationality are indeed self-confirming. In fact, we can show that this is true for every dynamic game! That is, in every dynamic game, the combination of rationality orderings induced by common strong belief in rationality is always self-confirming. Not only this, we can even show that it is the *only* self-confirming combination of rationality orderings in a game! Here is the reason why.

Take an arbitrary dynamic game, and for every player i a rationality ordering $R_i = (D_i^0, D_i^1, \dots, D_i^K)$ on his strategies, with the same K for all players. Suppose that this combination of rationality orderings is self-confirming. Then, by definition, the sets D_i^0 of 0-rational strategies should contain precisely those strategies that are not optimal under *any* conditional belief vector.

Turn now to the sets D_i^1 . Since the combination of rationality orderings is self-confirming, D_i^1 contains precisely those strategies that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level 0, but that are not optimal under any belief vector that strongly believes in the opponents' rationality orderings up to level 1. Now, the strategies that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level 0 are simply the strategies that are optimal under *some* conditional belief vector, without any further restrictions.

But which strategies are optimal under a belief vector that strongly believes in the opponents' rationality orderings up to level 1? Strong belief in the opponents' rationality orderings up to level 1 means for player

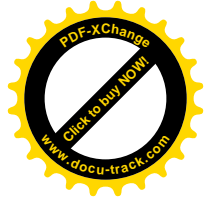


i that at every information set $h \in H_i$ he must see whether there is an opponents' combination of at-least-1-rational strategies leading to h . If so, then he must assign at h positive probability only to at-least-1-rational strategies for the opponents. Since for every opponent j , the set D_j^0 of 0-rational strategies contains precisely the strategies that are not optimal for any conditional belief vector, it follows that for every opponent j the at-least-1-rational strategies must be precisely those strategies that are optimal for some belief vector. That is, strong belief in the opponents' rationality orderings up to level 1 means for player i that at every information set $h \in H_i$ he must see whether there is an opponents' strategy combination leading to h where every strategy is optimal for some belief vector. If so, then he must assign at h positive probability only to strategies for the opponents that are optimal for some belief vector. This, however, is exactly the condition of 1-fold strong belief in rationality! Thus, the strategies that are optimal under a belief vector that strongly believes in the opponents' rationality orderings up to level 1, are precisely the strategies that can rationally be chosen under 1-fold strong belief in rationality.

Remember that D_i^1 contains precisely those strategies that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level 0, but that are not optimal under any belief vector that strongly believes in the opponents' rationality orderings up to level 1. But then, the set D_i^1 of 1-rational strategies contains precisely those strategies that can rationally be chosen under some conditional belief vector, but not under 1-fold strong belief in rationality. So, these are exactly the 1-rational strategies induced by common strong belief in rationality!

Next, consider the sets D_i^2 . As the combination of rationality orderings is self-confirming, D_i^2 contains precisely those strategies that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level 1, but that are not optimal under any belief vector that strongly believes in the opponents' rationality orderings up to level 2. We have seen above that the strategies that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level 1 are precisely those strategies that can rationally be chosen under 1-fold strong belief in rationality.

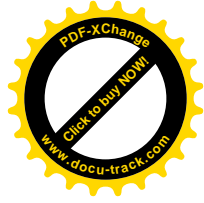
But which strategies can rationally be chosen under strong belief in the opponents' rationality orderings up to level 2? For player i , the condition of strong belief in the opponents' rationality orderings up to level 2 means that at every information set $h \in H_i$ he must see whether there is a strategy combination leading to h consisting of at-least-2-rational



strategies. If so, he must at h only assign positive probability to opponents' strategies that are at-least-2-rational. We have seen above that the set D_j^0 of 0-rational strategies contains those strategies that are not optimal for any beliefs, and the set D_j^1 of 1-rational strategies contains those strategies that are optimal under some conditional belief vector, but not under 1-fold strong belief in rationality. But then, the at-least-2-rational strategies for opponent j are precisely those strategies that are optimal under 1-fold strong belief in rationality. Hence, for player i the condition of strong belief in the opponents' rationality orderings up to level 2 means that at every information set $h \in H_i$ he must see whether there is an opponents' strategy combination leading to h where every strategy is optimal under 1-fold strong belief in rationality. If so, he must at h only assign positive probability to opponents' strategies that are optimal under 1-fold strong belief in rationality. But this is precisely the condition of 2-fold strong belief in rationality! Moreover, strong belief in the opponents' rationality orderings up to level 2 also means that at information set $h \in H_i$, player i must see whether there is an opponents' combination of at-least-1-rational strategies leading to h . If so, then he must assign at h positive probability only to at-least-1-rational strategies for the opponents. This, as we have seen, is equivalent to 1-fold strong belief in rationality. That is, the condition of strong belief in the opponents' rationality orderings up to level 2 is equivalent to expressing up to 2-fold strong belief in rationality. So, the strategies that can rationally be chosen under strong belief in the opponents' rationality orderings up to level 2 are precisely those strategies that can rationally be chosen under expressing up to 2-fold strong belief in rationality.

Remember that D_i^2 contains precisely those strategies that are optimal under a conditional belief vector that strongly believes in the opponents' rationality orderings up to level 1, but that are not optimal under any belief vector that strongly believes in the opponents' rationality orderings up to level 2. Hence, D_i^2 contains exactly those strategies that can rationally be chosen under 1-fold strong belief in rationality, but not under expressing up to 2-fold strong belief in rationality. So, these are exactly the 2-rational strategies induced by common strong belief in rationality.

By repeatedly using the arguments above, we conclude that for every k between 1 and $K - 1$, the sets D_i^k of k -rational strategies contain precisely those strategies that can rationally be chosen by expressing up to $(k - 1)$ -fold strong belief in rationality, but not under expressing up to k -fold strong belief in rationality. That is, these sets D_i^k would



be precisely the sets of k -rational strategies induced by common strong belief in rationality.

Let us finally consider the sets D_i^K of “most rational” strategies. As the combination of rationality orderings is self-confirming, D_i^K contains exactly those strategies that are optimal under strong belief in the opponents' rationality orderings up to level K . In particular, all strategies in D_i^K are optimal under strong belief in the opponents' rationality orderings up to level $K - 1$. By using a similar argument as above, it can be shown that all these strategies are optimal under expressing up to $(K - 1)$ -fold strong belief in rationality.

Moreover, all strategies in D_i^K are also optimal under strong belief in the opponents' rationality orderings up to level K . Here, strong belief in the opponents' rationality orderings up to level K means in particular that player i , at every information set $h \in H_i$, sees whether there is an opponents' strategy combination leading to h consisting of K -rational strategies – that is, strategies in D_j^K . If so, he must at h only assign positive probability to opponents' strategies s_j that are in D_j^K . We have seen above that all opponents' strategies in D_j^K are optimal under expressing up to $(K - 1)$ -fold strong belief in rationality. Hence, strong belief in the opponents' rationality orderings up to level K means that player i , at every information set $h \in H_i$, sees whether there is an opponents' strategy combination leading to h in which every strategy is optimal under expressing up to $(K - 1)$ -fold strong belief in rationality. If so, he must at h only assign positive probability to opponents' strategies that are optimal under expressing up to $(K - 1)$ -fold strong belief in rationality. This, however, is exactly the condition of K -fold strong belief in rationality. That is, strong belief in the opponents' rationality orderings up to level K implies expressing up to K -fold strong belief in rationality. We have seen that all strategies in D_i^K are optimal under strong belief in the opponents' rationality orderings up to level K . But then, it follows that all strategies in D_i^K can rationally be chosen under expressing up to K -fold strong belief in rationality.

So, not only can all strategies in D_i^K rationally be chosen under expressing up to $(K - 1)$ -fold strong belief in rationality as we have seen above, but they can also rationally be chosen under expressing up to K -fold strong belief in rationality. By basically repeating the argument above, we can then show that all strategies in D_i^K can also rationally be chosen under expressing up to $(K + 1)$ -fold strong belief in rationality, under expressing up to $(K + 2)$ -fold strong belief in rationality, and so on. This, however, means that all strategies in D_i^K can rationally be chosen under common strong belief in rationality!



Summarizing, we see

- that the sets D_i^0 contain exactly those strategies that are not optimal under any belief vector,
- that for every k between 1 and $K - 1$, the sets D_i^k contain precisely those strategies that can rationally be chosen under expressing up to $(k - 1)$ -fold strong belief in rationality, but not under expressing up to k -fold strong belief in rationality, and
- that the sets D_i^K contain precisely those strategies that can rationally be chosen under common strong belief in rationality.

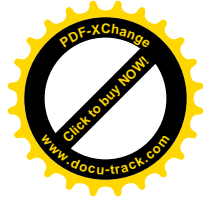
This means, however, that the rationality orderings

$R_i = (D_i^0, D_i^1, \dots, D_i^K)$ we started with must exactly be the rationality orderings induced by common strong belief in rationality. So, we have shown that there is only one combination of rationality orderings that is self-confirming, namely the combination of rationality orderings induced by common strong belief in rationality. We have thus established the following general result.

THEOREM 9.6.5. (*Characterization of self-confirming combinations of rationality orderings*)

Consider an arbitrary dynamic game, and let K be the number of steps after which the iterated conditional dominance procedure stops. Then, there is only one self-confirming combination of rationality orderings, namely that in which each rationality ordering R_i is given by $R_i = (D_i^0, D_i^1, \dots, D_i^K)$, where (1) the sets D_i^0 contain exactly those strategies that are not optimal under any belief vector, (2) for every k between 1 and $K - 1$, the sets D_i^k contain precisely those strategies that can rationally be chosen under expressing up to $(k - 1)$ -fold strong belief in rationality, but not under expressing up to k -fold strong belief in rationality, and (3) the sets D_i^K contain precisely those strategies that can rationally be chosen under common strong belief in rationality.

We know that for every k , the strategies that can rationally be chosen under expressing up to k -fold strong belief in rationality are exactly the strategies in $\Gamma^{k+1}(\emptyset)$ in the iterated conditional dominance procedure. So, the unique self-confirming combination of rationality orderings will always be such that (1) for every k between 0 and $K - 1$, the sets D_i^k contain precisely those strategies that are in $\Gamma^k(\emptyset)$ but not in $\Gamma^{k+1}(\emptyset)$, and (2) the sets D_i^K contain exactly the strategies in $\Gamma^K(\emptyset)$, where K is the number of steps after which the algorithm stops. That is, the iterated conditional dominance procedure provides a fast and easy method to generate the unique self-confirming combination of rationality orderings in a dynamic game.

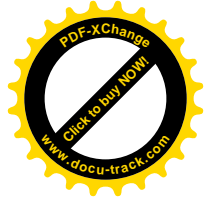


On a conceptual level, the theorem above shows that common strong belief in rationality can be characterized by the use of rationality orderings on strategies. More precisely, a player who expresses common strong belief in rationality has in his mind a rationality ordering for each of his opponents, and uses these rationality orderings to form his beliefs at each of his information sets. So, at every information set h he looks for the highest degree of rationality for his opponents – as given by these rationality orderings – that makes reaching h possible, and uses this highest possible degree of rationality to form his beliefs at h . This is called the *best rationalization principle*. In particular, a player who expresses common strong belief in rationality always uses the *same* rationality orderings on opponents' strategies throughout the game.

Let us compare this to the concept of common belief in future rationality which we discussed in the previous chapter. Is there a way to characterize also this concept in terms of rationality orderings and the best rationalization principle above? We will see that this is not possible.

To see this, let us consider the example “Watching TV with Barbara” that we know so well by now. It turns out that under common belief in future rationality, you can rationally choose your strategies $(fight, B)$, $(don't, B)$ and $(don't, D)$, whereas Barbara can rationally choose any of her strategies under this concept. Suppose now that Barbara expresses common belief in future rationality, and that her beliefs would be induced by a rationality ordering on your strategies. Since each of your strategies $(fight, B)$, $(don't, B)$ and $(don't, D)$ are optimal under common belief in future rationality, the “most rational” strategies for you in this ordering must be $\{(fight, B), (don't, B), (don't, D)\}$ whereas the “least rational” strategy for you must be $(fight, D)$. But then, at information set h_1 – that is, after you have chosen “fight” – Barbara would see that there is only one “most rational” strategy for you leading to h_1 , which is $(fight, B)$. So, if her beliefs are induced by this rationality ordering, she must believe at h_1 that you have chosen $(fight, B)$. But then, Barbara can no longer rationally choose (D, B) or (D, D) , which is a contradiction since the latter two strategies can rationally be chosen under common belief in future rationality. This shows that common belief in future rationality cannot be characterized by rationality orderings on strategies.

The intuitive reason for this is that under common belief in future rationality, Barbara's ordering of your strategies may actually *change* once the game is under way. Initially, that is, before the game starts, Barbara deems your strategies $(fight, B)$, $(don't, B)$ and $(don't, D)$ indeed “more rational” than your strategy $(fight, D)$. But at information set h_1 , after you have started a fight with her, she does not necessarily deem



your strategy $(fight, B)$ “more rational” than your strategy $(fight, D)$, as both strategies can still be optimal for you *at information set* h_1 under common belief in future rationality. Remember that under common belief in future rationality, players do not need to believe that opponents have chosen rationally in the past, only that they choose rationally now and in the future. Hence, we may conclude that rationality orderings are really something typical for common strong belief in rationality, and cannot be applied for common belief in future rationality.

9.7. Bayesian Updating

In this chapter we have presented, and explored, the concept of *common strong belief in rationality*. As in Chapter 8, we can ask what would happen to the concept if we would require, in addition, that players satisfy *Bayesian updating* when revising their beliefs. That is, suppose we would restrict the concept of common strong belief in rationality by additionally imposing common belief in Bayesian updating. What can we say about the strategies that can rationally be chosen under this more restrictive concept?

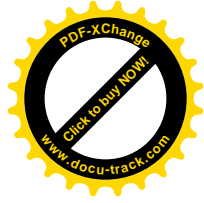
We have seen in Chapter 8 that for the concept of *common belief in future rationality*, it crucially matters whether we impose (common belief in) Bayesian updating or not. There are games, namely, in which some strategy can rationally be chosen under common belief in future rationality, but where the same strategy can no longer be rationally chosen if we additionally impose common belief in Bayesian updating. See the example of Figure 8.14 in that chapter.

For the concept of common strong belief in rationality, the story is different. It can be shown, namely, that the sets of strategies that can rationally be chosen under common strong belief in rationality would remain the same if we additionally were to impose common belief in Bayesian updating. That is, additionally requiring common belief in Bayesian updating has no consequences for the eventual strategy choices selected by the concept. We thus obtain the following result.

THEOREM 9.7.1. (*Bayesian updating is irrelevant for common strong belief in rationality*)

Every strategy that can rationally be chosen under common strong belief in rationality, can also rationally be chosen under common strong belief in rationality and common belief in Bayesian updating.

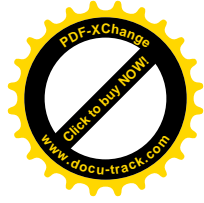
Here, we say that a type t_i expresses *common belief in Bayesian updating* if t_i assigns, at every $h \in H_i$, only positive probability to opponents' types that satisfy Bayesian updating, if t_i assigns, at every $h \in H_i$,



only positive probability to opponents' types t_j that, every $h' \in H_j$, only assign positive probability to opponents' types that satisfy Bayesian updating, and so on. Moreover, we say that a strategy s_i can rationally be chosen under common strong belief in rationality and common belief in Bayesian updating if s_i is optimal for some type t_i that expresses common strong belief in rationality, satisfies Bayesian updating and expresses common belief in Bayesian updating.

The idea for proving Theorem 9.7.1 is the following. In Theorem 9.3.3 we have shown that the strategies that can rationally be chosen under common strong belief in rationality, are exactly the strategies that survive the iterated conditional dominance procedure at \emptyset . In the proof of Theorem 9.3.3 we construct an epistemic model M in which, for each of the strategies s_i that survive the iterated conditional dominance procedure at \emptyset , there is some type $t_i^{s_i}$ that expresses common strong belief in rationality, and for which s_i is optimal. In this epistemic model, we could as well have constructed these types $t_i^{s_i}$ in such a way that they satisfy Bayesian updating and express common belief in Bayesian updating. We leave it to the interested reader to adapt the proof in this way. But then, we conclude that every strategy s_i which survives the iterated conditional dominance procedure at \emptyset , is optimal for some type $t_i^{s_i}$ which expresses common strong belief in rationality, satisfies Bayesian updating, and expresses common belief in Bayesian updating.

That is, every strategy which survives the iterated conditional dominance procedure at \emptyset , can rationally be chosen under common strong belief in rationality and common belief in Bayesian updating. By Theorem 9.3.3 it then follows that the strategies that can rationally be chosen under common strong belief in rationality are the same as the strategies that can rationally be chosen under common strong belief in rationality and common belief in Bayesian updating. In both cases, namely, we obtain exactly those strategies that survive the iterated conditional dominance procedure at \emptyset .



9.8. Proofs

In this section we will first prove Theorem 9.3.3, which states that the iterated conditional dominance procedure delivers exactly those strategies that can rationally be chosen under common strong belief in rationality. We proceed in three steps. We first derive some important properties of the iterated conditional dominance procedure. In the second subsection we use these properties to prove an *optimality principle* for the iterated conditional dominance procedure. As you will see, proving this optimality principle will be far from easy. In the third subsection we prove Theorem 9.3.3, heavily relying on the optimality principle we derive in the second subsection.

In the fourth and final subsection we will prove Theorem 9.4.2, which states that every outcome you initially deem possible under common strong belief in rationality, is also initially deemed possible under common belief in future rationality.

Properties of the Iterated Conditional Dominance Procedure. In this first subsection we will prove some important properties of the iterated conditional dominance procedure, which we will use in the following subsection to show the optimality principle. Recall that $\Gamma^k(\emptyset)$ is the decision problem that remains after step k of the iterated conditional dominance procedure at \emptyset . For every player i and every information set h , let $\Gamma^k(h) = (S_i^k(h), S_{-i}^k(h))$ be the decision problem that remains after step k of the iterated conditional dominance procedure at h . We first show how $\Gamma^k(h)$ relates to the decision problems $\Gamma^m(\emptyset)$ for $m \leq k$.

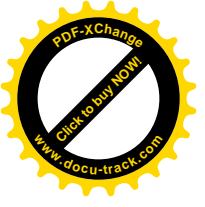
LEMMA 9.8.1. (*Structure of the decision problems $\Gamma^k(h)$*)

For some player i and some information set $h \in H_i$, let $\Gamma^k(h) = (S_i^k(h), S_{-i}^k(h))$ be the decision problem that remains after step k of the iterated conditional dominance procedure at h . Let m be the highest number less than or equal to k such that some strategy combination in $\Gamma^m(\emptyset)$ leads to h . Then, $S_i^k(h) = S_i^m(\emptyset) \cap S_i(h)$, and $S_{-i}^k(h) = S_{-i}^m(\emptyset) \cap S_{-i}(h)$.

Remember that $S_i(h)$ contains exactly those strategies for player i that lead to h , and $S_{-i}(h)$ contains exactly those opponents' strategy combinations that lead to h .

Proof. We prove the statement by induction on k .

Induction start. Start with $k = 0$. By definition, $S_i^0(h) = S_i(h)$ and $S_{-i}^0(h) = S_{-i}(h)$. Moreover, $S_i^0(\emptyset) = S_i$ and $S_{-i}^0(\emptyset) = S_{-i}$. So, the statement trivially follows for $k = m = 0$.



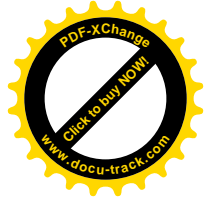
Induction step. Take some $k \geq 1$, and assume that the statement is true for every $k' \leq k - 1$. Take some information set $h \in H_i$, and let m be the highest number less than or equal to k such that some strategy combination in $\Gamma^m(\emptyset)$ leads to h . We distinguish two cases.

Case 1. If $m < k$, then there is no strategy combination in $\Gamma^k(\emptyset)$ that leads to h . So, every strategy combination that leads to h contains at least one strategy which is not in $\Gamma^k(\emptyset)$. By construction, $\Gamma^k(\emptyset)$ contains precisely those strategies s_j that are not strictly dominated at any decision problem $\Gamma^{k-1}(h')$ of step $k - 1$ at which j is active. Hence, every strategy combination that leads to h contains at least one strategy s_j which is strictly dominated in some decision problem $\Gamma^{k-1}(h')$ of step $k - 1$ at which j is active. By definition of the iterated conditional dominance procedure, we should then eliminate no strategies from $\Gamma^{k-1}(h)$ in step k . That is, $\Gamma^k(h)$ is the same as $\Gamma^{k-1}(h)$. But then, m is clearly the highest number less than or equal to $k - 1$ such that some strategy combination in $\Gamma^m(\emptyset)$ leads to h . So, we know by our induction assumption that $\Gamma^{k-1}(h)$ contains precisely those strategy combinations in $\Gamma^m(\emptyset)$ that lead to h . In other words, $S_i^{k-1}(h) = S_i^m(\emptyset) \cap S_i(h)$, and $S_{-i}^{k-1}(h) = S_{-i}^m(\emptyset) \cap S_{-i}(h)$. As $\Gamma^{k-1}(h) = \Gamma^k(h)$, the same holds for $\Gamma^k(h)$. Hence, $S_i^k(h) = S_i^m(\emptyset) \cap S_i(h)$, and $S_{-i}^k(h) = S_{-i}^m(\emptyset) \cap S_{-i}(h)$.

Case 2. Suppose that $m = k$. So, there is a strategy combination in $\Gamma^k(\emptyset)$ that leads to h . We show that $\Gamma^k(h)$ contains precisely those strategy combinations in $\Gamma^k(\emptyset)$ that lead to h .

By definition of the iterated conditional dominance procedure, $\Gamma^k(\emptyset)$ contains precisely those strategies s_j that are not strictly dominated in any decision problem $\Gamma^{k-1}(h')$ of step $k - 1$ at which j is active. By our assumption, at least one such strategy combination leads to h . Hence, there is a combination of strategies s_j leading to h (including i 's strategy) where every strategy s_j is not strictly dominated in any decision problem $\Gamma^{k-1}(h')$ of step $k - 1$ at which j is active. By definition of the iterated conditional dominance procedure, $\Gamma^k(h)$ then contains precisely those strategy combinations leading to h where every strategy s_j is not strictly dominated in any decision problem $\Gamma^{k-1}(h')$ of step $k - 1$ at which j is active. So, $\Gamma^k(h)$ contains precisely those strategy combinations in $\Gamma^k(\emptyset)$ that lead to h . That is, $S_i^k(h) = S_i^k(\emptyset) \cap S_i(h)$, and $S_{-i}^k(h) = S_{-i}^k(\emptyset) \cap S_{-i}(h)$, which was to show.

We may thus conclude that, in general, $S_i^k(h) = S_i^m(\emptyset) \cap S_i(h)$, and $S_{-i}^k(h) = S_{-i}^m(\emptyset) \cap S_{-i}(h)$. By induction on k , the statement in the lemma follows. \diamond



In the proofs section of Chapter 8 we have introduced the so-called “forward inclusion property”. We repeat its definition here for convenience. For a given player i , take a collection $(D_{-i}(h))_{h \in H_i}$ of strategy subsets, specifying at every information set $h \in H_i$ some subset $D_{-i}(h) \subseteq S_{-i}(h)$ of opponents’ strategy combinations leading to h . The collection $(D_{-i}(h))_{h \in H_i}$ of strategy subsets is said to satisfy the *forward inclusion property* if for every $h, h' \in H_i$ where h' follows h , it holds that $D_{-i}(h) \cap S_{-i}(h') \subseteq D_{-i}(h')$. We will now show that the collections of strategy subsets selected by the iterated conditional dominance procedure satisfy the forward inclusion property.

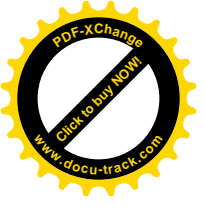
LEMMA 9.8.2. (*Iterated conditional dominance procedure satisfies forward inclusion property*)

For every player i and information set $h \in H_i$, let $\Gamma^k(h) = (S_i^k(h), S_{-i}^k(h))$ be the decision problem that remains after step k of the iterated conditional dominance procedure at h . Then, the collection $(S_{-i}^k(h))_{h \in H_i}$ of strategy subsets satisfies the forward inclusion property.

Proof. Take some information sets $h, h' \in H_i$ where h' follows h . We must show that $S_{-i}^k(h) \cap S_{-i}(h') \subseteq S_{-i}^k(h')$. Let m be the highest number less than or equal to k such that there is a strategy combination in $\Gamma^m(\emptyset)$ leading to h . Then, by Lemma 9.8.1 we know that $S_{-i}^k(h) = S_{-i}^m(\emptyset) \cap S_{-i}(h)$. Similarly, let m' be the highest number less than or equal to k such that there is a strategy combination in $\Gamma^{m'}(\emptyset)$ leading to h' . Since h' follows h we have that $m' \leq m$, and by Lemma 9.8.1 we know that $S_{-i}^k(h') = S_{-i}^{m'}(\emptyset) \cap S_{-i}(h')$.

Since $S_{-i}^k(h) = S_{-i}^m(\emptyset) \cap S_{-i}(h)$, and $S_{-i}(h') \subseteq S_{-i}(h)$, we have that $S_{-i}^k(h) \cap S_{-i}(h') = S_{-i}^m(\emptyset) \cap S_{-i}(h')$. Moreover, $S_{-i}^m(\emptyset) \subseteq S_{-i}^{m'}(\emptyset)$ as $m \geq m'$. Hence, we conclude that $S_{-i}^k(h) \cap S_{-i}(h') \subseteq S_{-i}^{m'}(\emptyset) \cap S_{-i}(h') = S_{-i}^k(h')$, which was to show. We thus have proved that $(S_{-i}^k(h))_{h \in H_i}$ satisfies the forward inclusion property. \diamond

Optimality Principle. Recall that for every player i and information set $h \in H_i$, we denote by $\Gamma^k(h) = (S_i^k(h), S_{-i}^k(h))$ the reduced decision problem for player i at h after applying step k of the iterated conditional dominance procedure. Moreover, we denote by $S_i^k(\emptyset)$ the set of strategies for player i that survive step k of the iterated conditional dominance procedure at \emptyset . By definition of the iterated conditional dominance procedure, $S_i^k(\emptyset)$ contains precisely those strategies s_i for player i that are not strictly dominated within any decision problem $\Gamma^{k-1}(h)$ for player i that s_i leads to. For every such information set h , the decision problem $\Gamma^{k-1}(h)$ is given by $(S_i^{k-1}(h), S_{-i}^{k-1}(h))$. By Theorem 2.5.3 in



Chapter 2 we know that a strategy s_i is not strictly dominated in $\Gamma^{k-1}(h)$ precisely when it is optimal, among strategies in $S_i^{k-1}(h)$, for some belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$. That is,

$$u_i(s_i, b_i(h)) \geq u_i(s'_i, b_i(h)) \text{ for all } s'_i \in S_i^{k-1}(h).$$

Hence, $S_i^k(\emptyset)$ contains precisely those strategies s_i for player i which, at every $h \in H_i$ that s_i leads to, are optimal for some belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ among all strategies in $S_i^{k-1}(h)$.

We can show, however, that these strategies s_i are not only optimal at h among strategies in $S_i^{k-1}(h)$ only, but in fact among all strategies in $S_i(h)$. That is,

$$u_i(s_i, b_i(h)) \geq u_i(s'_i, b_i(h)) \text{ for all } s'_i \in S_i(h).$$

We will refer to this as the *optimality principle* for the iterated conditional dominance procedure. Like in Chapter 8, it will play an important role in proving that the iterated conditional dominance procedure yields exactly those strategies that can rationally be chosen under common strong belief in rationality.

LEMMA 9.8.3. (*Optimality principle*)

For some $k \geq 1$, consider a strategy $s_i \in S_i^k(\emptyset)$ that survives step k of the iterated conditional dominance procedure at \emptyset . Then, at every $h \in H_i$ that s_i leads to, there is some belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ such that s_i is optimal for $b_i(h)$ among all strategies in $S_i(h)$.

Proof. Consider a strategy $s_i^* \in S_i^k(\emptyset)$ and an information set $h^* \in H_i$ that s_i^* leads to. We show that there is some belief $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$ such that s_i^* is optimal for $b_i(h^*)$ among all strategies in $S_i(h^*)$. We prove this statement by induction on the number of player i information sets that precede h^* .

Induction start. Take an information set $h^* \in H_i$ that s_i^* leads to, which is not preceded by any other player i information set. As $s_i^* \in S_i^k(\emptyset)$, we know by the argument above that there is some belief $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$ such that s_i^* is optimal for $b_i(h^*)$ among all strategies in $S_i^{k-1}(h^*)$. That is,

$$(1) \quad u_i(s_i^*, b_i(h^*)) \geq u_i(s'_i, b_i(h^*)) \text{ for all } s'_i \in S_i^{k-1}(h^*).$$

We will prove that, in fact,

$$u_i(s_i^*, b_i(h^*)) \geq u_i(s'_i, b_i(h^*)) \text{ for all } s'_i \in S_i(h^*).$$

Suppose, on the contrary, that there would be some $s'_i \in S_i(h^*)$ such that

$$(2) \quad u_i(s_i^*, b_i(h^*)) < u_i(s'_i, b_i(h^*)).$$



We show that in this case there would be some $s_i \in S_i^{k-1}(h^*)$ with $u_i(s'_i, b_i(h^*)) \leq u_i(s_i, b_i(h^*))$, which together with (2) would contradict (1).

Remember from Lemma 9.8.2 that the collection of strategy subsets $(S_{-i}^{k-1}(h))_{h \in H_i}$ satisfies the forward inclusion property. Since $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$, we know from Lemma 8.14.3 in the proofs section of Chapter 8 that we can choose at every information set $h \in H_i$ following h^* some belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ such that these beliefs, together with $b_i(h^*)$, satisfy Bayesian updating at all player i information sets following h^* .

Let H_i^{first} be the collection of player i information sets that are not preceded by any other player i information set. For every $h \in H_i^{first}$ other than h^* , choose some arbitrary belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$. Then, by the same argument as above, we can choose at every information set $h' \in H_i$ following h some belief $b_i(h') \in \Delta(S_{-i}^{k-1}(h'))$ such that these beliefs, together with $b_i(h)$, satisfy Bayesian updating at all player i information sets following h .

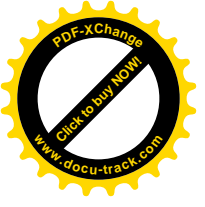
So, in this way we can extend the belief $b_i(h^*)$ to a conditional belief vector $(b_i(h))_{h \in H_i}$ on $(S_{-i}^{k-1}(h))_{h \in H_i}$ which satisfies Bayesian updating everywhere. But then, by Lemma 8.13.2 of Chapter 8, we know that there is a strategy s_i which, at every $h \in H_i$ that s_i leads to, is optimal for $b_i(h)$ among all strategies in $S_i(h)$. Hence, at every $h \in H_i$ that s_i leads to, strategy s_i is not strictly dominated on $S_{-i}^{k-1}(h)$ among all strategies in $S_i(h)$. So, $s_i \in S_i^k(\emptyset)$, and in particular $s_i \in S_i^{k-1}(\emptyset)$. Moreover, as there is no player i information set preceding h^* , it must be true that $s_i \in S_i(h^*)$. Hence, $s_i \in S_i^{k-1}(\emptyset) \cap S_i(h^*)$. By Lemma 9.8.1 we know that $S_i^{k-1}(\emptyset) \cap S_i(h^*) \subseteq S_i^{k-1}(h^*)$, and therefore we conclude that $s_i \in S_i^{k-1}(h^*)$.

By construction, strategy s_i is optimal at h^* , among all strategies in $S_i(h^*)$, for the belief $b_i(h^*)$. But then, $u_i(s'_i, b_i(h^*)) \leq u_i(s_i, b_i(h^*))$. Together with (2) this would imply that

$$u_i(s_i^*, b_i(h^*)) < u_i(s'_i, b_i(h^*)) \leq u_i(s_i, b_i(h^*)) \text{ for some } s_i \in S_i^{k-1}(h^*).$$

This, however, contradicts (1). So, strategy s_i^* must be optimal for the belief $b_i(h^*)$ among all strategies in $S_i(h^*)$, which is what we wanted to show.

Induction step. Consider some information set $h^* \in H_i$ that s_i^* leads to, and assume that for every $h \in H_i$ preceding h^* there is some belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ such that s_i^* is optimal for $b_i(h)$ among all strategies in $S_i(h)$. We prove that there is a belief $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$ such that



s_i^* is optimal for $b_i(h^*)$ among all strategies in $S_i(h^*)$. To do so, we distinguish two cases.

Case 1. Suppose that there is a player i information set h preceding h^* with $b_i(h)(S_{-i}(h^*)) > 0$. By our induction assumption, we know that s_i^* is optimal at h for the belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ among all strategies in $S_i(h)$. Let $b_i(h^*)$ be the belief at h^* obtained from $b_i(h)$ by Bayesian updating. But then, by Lemma 8.14.9 from the proofs section in Chapter 8, it follows that s_i^* is also optimal at h^* for the belief $b_i(h^*)$ among all strategies in $S_i(h^*)$.

We will now show that $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$. As $b_i(h^*)$ is obtained from $b_i(h)$ by Bayesian updating, we know that

$$(1) \quad b_i(h^*)(s_{-i}) = \frac{b_i(h)(s_{-i})}{b_i(h)(S_{-i}(h^*))}$$

for every opponents' strategy combination $s_{-i} \in S_{-i}(h^*)$. By assumption, $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$, and hence $b_i(h)$ only assigns positive probability to opponents' strategy combinations in $S_{-i}^{k-1}(h)$. But then, it follows by (1) that $b_i(h^*)$ only assigns positive probability to opponents' strategy combinations in $S_{-i}^{k-1}(h) \cap S_{-i}(h^*)$. From Lemma 9.8.2 we know that the collection of strategy subsets $(S_{-i}^{k-1}(h'))_{h' \in H_i}$ satisfies the forward inclusion property, and hence $S_{-i}^{k-1}(h) \cap S_{-i}(h^*) \subseteq S_{-i}^{k-1}(h^*)$. So, we conclude that $b_i(h^*)$ only assigns positive probability to opponents' strategy combinations in $S_{-i}^{k-1}(h^*)$, that is, $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$.

Hence, we have constructed a belief $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$ such that s_i^* is optimal for $b_i(h^*)$ among all strategies in $S_i(h^*)$. This completes Case 1.

Case 2. Suppose that $b_i(h)(S_{-i}(h^*)) = 0$ for every player i information set h preceding h^* .

Since $s_i^* \in S_i^k(\emptyset)$ we know from above that at h^* there is some belief $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$ such that s_i^* is optimal for $b_i(h^*)$ among all strategies in $S_i^{k-1}(h^*)$. That is,

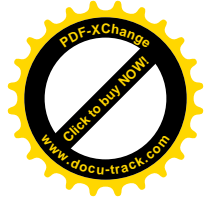
$$(2) \quad u_i(s_i^*, b_i(h^*)) \geq u_i(s'_i, b_i(h^*)) \text{ for all } s'_i \in S_i^{k-1}(h^*).$$

We will prove that, in fact,

$$u_i(s_i^*, b_i(h^*)) \geq u_i(s'_i, b_i(h^*)) \text{ for all } s'_i \in S_i(h^*).$$

Suppose, on the contrary, that there would be some $s'_i \in S_i(h^*)$ such that

$$(3) \quad u_i(s_i^*, b_i(h^*)) < u_i(s'_i, b_i(h^*)).$$

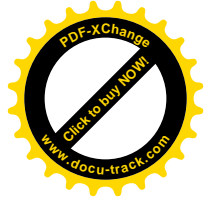


We show that in this case there would be some $s_i \in S_i^{k-1}(h^*)$ with $u_i(s'_i, b_i(h^*)) \leq u_i(s_i, b_i(h^*))$, which together with (3) would contradict (2).

Denote by $H_i^{pre}(h^*)$ the collection of player i information sets that precede h^* . Let H_i^* be the collection of player i information sets $h \notin H_i^{pre}(h^*)$ such that (1) h follows some information set $h' \in H_i^{pre}(h^*)$, and (2) $b_i(h')(S_{-i}(h)) > 0$ for this information set $h' \in H_i^{pre}(h^*)$. Remember that, by our assumption in Case 2, $b_i(h')(S_{-i}(h^*)) = 0$ for every $h' \in H_i^{pre}(h^*)$, and hence $h^* \notin H_i^*$. Take some information set $h \in H_i^*$, preceded by some $h' \in H_i^{pre}(h^*)$ with $b_i(h')(S_{-i}(h)) > 0$. Define the belief $b_i(h)$ at h to be the belief obtained from $b_i(h')$ by Bayesian updating. Since $b_i(h') \in \Delta(S_{-i}^{k-1}(h'))$, it can be shown in the same way as above that $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$. Since $h' \in H_i^{pre}(h^*)$, we know by our induction assumption that s_i^* is optimal at h' , among all strategies in $S_i(h')$, for the belief $b_i(h')$. But then, it follows from Lemma 8.14.9 from the proofs section in Chapter 8 that s_i^* is also optimal at h for $b_i(h)$ among all strategies in $S_i(h)$, if s_i^* leads to h . So, we see that for every $h \in H_i^*$ there is a belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ such that s_i^* is optimal at h for $b_i(h)$ among all strategies in $S_i(h)$, if s_i^* leads to h .

Denote by H_i^0 the collection of player i information sets which are not in $H_i^{pre}(h^*)$ and not in H_i^* . So, H_i^0 contains those player i information sets h that do not precede h^* , and that are either not preceded by any information set $h' \in H_i^{pre}(h^*)$, or for which $b_i(h')(S_{-i}(h)) = 0$ for every $h' \in H_i^{pre}(h^*)$ that precedes h . Note that, by our assumption for Case 2, $b_i(h')(S_{-i}(h^*)) = 0$ for every $h' \in H_i^{pre}(h^*)$, and hence $h^* \in H_i^0$. By construction, if $h \in H_i^0$ and $h' \in H_i$ follows h , then also h' is in H_i^0 . By H_i^{0first} we denote the collection of those information sets $h \in H_i^0$ that are not preceded by any other information set in H_i^0 . As $h^* \in H_i^0$, and every player i information set preceding h^* is in $H_i^{pre}(h^*)$, we conclude that $h^* \in H_i^{0first}$.

At information set $h^* \in H_i^{0first}$, we had already defined the belief $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$ as above, satisfying (2). At every other information set $h \in H_i^{0first}$ we choose some arbitrary belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$. Now, take some arbitrary information set $h \in H_i^{0first}$. As the collection of strategy subsets $(S_{-i}^{k-1}(h'))_{h' \in H_i}$ satisfies the forward inclusion property, we know from Lemma 8.14.3 in the proofs section of Chapter 8 that we can find beliefs $b_i(h') \in \Delta(S_{-i}^{k-1}(h'))$ at all $h' \in H_i$ following h such that these beliefs, together with $b_i(h)$, satisfy Bayesian updating at all $h' \in H_i$ following h . But then, by Lemma 8.13.2 from Chapter 8, we can find a strategy $s_i^h \in S_i(h)$ such that, at every $h' \in H_i$ weakly following



h that s_i^h leads to, s_i^h is optimal at h' for $b_i(h')$ among all strategies in $S_i(h')$. So, for every $h \in H_i^{0first}$ there is a strategy $s_i^h \in S_i(h)$, and beliefs $b_i(h') \in \Delta(S_{-i}^{k-1}(h'))$ at every $h' \in H_i$ weakly following h , such that, at every $h' \in H_i$ weakly following h that s_i^h leads to, s_i^h is optimal at h' for $b_i(h')$ among all strategies in $S_i(h')$.

We now construct a strategy s_i as follows. At every $h \in H_i^{pre}(h^*)$, and every $h \in H_i^*$ that s_i^* leads to, let s_i select the same choice as s_i^* . At every $h \in H_i^0$, weakly preceded by some $h' \in H_i^{0first}$, let s_i select the same choice as $s_i^{h'}$, if $s_i^{h'}$ leads to h . So, in short, strategy s_i coincides with s_i^* at information sets in $H_i^{pre}(h^*)$ and H_i^* , and, for every $h \in H_i^{0first}$, coincides with s_i^h at information sets that weakly follow h .

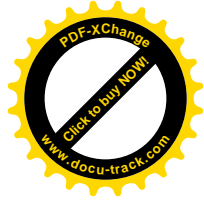
We will now show that, at every $h \in H_i$ that s_i leads to, the strategy s_i is optimal for the belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ among all strategies in $S_i(h)$. To do so, we distinguish three cases.

Case a. Suppose that $h \in H_i^{pre}(h^*)$. By construction, every information set $h' \in H_i$ following h with $b_i(h)(S_{-i}(h')) > 0$ is in H_i^* . That is, in order to verify the optimality of a strategy s_i'' for the belief $b_i(h)$ at h we only need the choices that s_i'' prescribes at h , and at information sets in H_i^* . By construction, the strategy s_i coincides with s_i^* at h , and at information sets in H_i^* . Moreover, by our induction assumption we know that strategy s_i^* is optimal at h for the belief $b_i(h)$ among all strategies in $S_i(h)$. Therefore, we conclude that also strategy s_i is optimal at h for the belief $b_i(h)$ among all strategies in $S_i(h)$.

Case b. Suppose that $h \in H_i^*$. We show that every information set $h' \in H_i$ following h with $b_i(h)(S_{-i}(h')) > 0$ is also in H_i^* . Namely, take some information set $h' \in H_i$ following h with $b_i(h)(S_{-i}(h')) > 0$. As $h \in H_i^*$, we know that h is preceded by some $h'' \in H_i^{pre}(h^*)$ with $b_i(h'')(S_{-i}(h)) > 0$, and that $b_i(h)$ is obtained from $b_i(h'')$ by Bayesian updating. As $b_i(h)(S_{-i}(h')) > 0$, and $b_i(h)$ is obtained from $b_i(h'')$ by Bayesian updating, it follows that also $b_i(h'')(S_{-i}(h')) > 0$. This, however, means that $h' \in H_i^*$, since $h'' \in H_i^{pre}(h^*)$. So, indeed, every information set $h' \in H_i$ following h with $b_i(h)(S_{-i}(h')) > 0$ is in H_i^* .

Hence, in order to verify the optimality of a strategy s_i'' for the belief $b_i(h)$ at h we only need the choices that s_i'' prescribes at information sets in H_i^* . By construction, s_i coincides with s_i^* at information sets in H_i^* . Moreover, we know by our construction that s_i^* is optimal at h for the belief $b_i(h)$ among all strategies in $S_i(h)$. We thus conclude that also strategy s_i is optimal at h for the belief $b_i(h)$ among all strategies in $S_i(h)$.

Case c. Suppose that $h \in H_i^0$. Let h be weakly preceded by some $h' \in H_i^{0first}$. By our construction, strategy s_i coincides with $s_i^{h'}$ at h



and all player i information sets following h . Moreover, we know that strategy $s_i^{h'}$ is optimal at h for $b_i(h)$ among all strategies in $S_i(h)$. Hence, we conclude that also strategy s_i is optimal at h for $b_i(h)$ among all strategies in $S_i(h)$.

In total, we see that, at every $h \in H_i$ that s_i leads to, the strategy s_i is optimal for the belief $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ among all strategies in $S_i(h)$. Hence, strategy s_i is not strictly dominated at any decision problem $\Gamma^{k-1}(h) = (S_i^{k-1}(h), S_{-i}^{k-1}(h))$ for player i that s_i leads to. This means, however, that $s_i \in S_i^k(\emptyset)$, and hence in particular $s_i \in S_i^{k-1}(\emptyset)$. Moreover, it is easy to see that strategy s_i leads to h^* . Namely, by our construction, strategy s_i coincides with s_i^* at all player i information sets that precede h^* . Since s_i^* leads to h^* , we may conclude that s_i leads to h^* as well. Hence, we see that $s_i \in S_i^{k-1}(\emptyset) \cap S_i(h^*)$. By Lemma 9.8.1 we know that $S_i^{k-1}(\emptyset) \cap S_i(h^*) \subseteq S_i^{k-1}(h^*)$, and hence $s_i \in S_i^{k-1}(h^*)$.

So, we have constructed a strategy $s_i \in S_i^{k-1}(h^*)$ which is optimal, at every $h \in H_i$ that s_i leads to, for the belief $b_i(h)$ among all strategies in $S_i(h)$. In particular, strategy s_i is optimal at h^* for the belief $b_i(h^*)$ among all strategies in $S_i(h^*)$. But then, by (3) it follows that

$$u_i(s_i^*, b_i(h^*)) < u_i(s_i', b_i(h^*)) \leq u_i(s_i, b_i(h^*)) \text{ for some } s_i \in S_i^{k-1}(h^*),$$

which contradicts (2). Hence, strategy s_i^* must be optimal at h^* for the belief $b_i(h^*)$ among all strategies in $S_i(h^*)$. So, we have constructed a belief $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$ such that s_i^* is optimal for $b_i(h^*)$ among all strategies in $S_i(h^*)$. This completes Case 2.

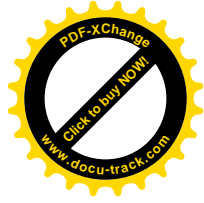
By induction, this statement will then hold for every information set $h^* \in H_i$ that s_i^* leads to. So, for every $h^* \in H_i$ that s_i^* leads to, we can construct some belief $b_i(h^*) \in \Delta(S_{-i}^{k-1}(h^*))$ such that s_i^* is optimal for $b_i(h^*)$ among all strategies in $S_i(h^*)$. This completes the proof of the optimality principle. \diamond

Algorithm “Works”. We will now use the optimality principle to prove the main theorem of this chapter.

THEOREM 9.3.3 (*Algorithm “works”*)

(1) For every $k \geq 1$, the strategies that can rationally be chosen by a type that expresses up to k -fold strong belief in rationality are precisely the strategies in $\Gamma^{k+1}(\emptyset)$ – that is, those strategies that survive step $k+1$ of the iterated conditional dominance procedure at \emptyset .

(2) The strategies that can rationally be chosen by a type that expresses common strong belief in rationality are exactly the strategies that survive the full iterated conditional dominance procedure at \emptyset .



Proof. For every $k \geq 1$, every player i , and every information set $h \in H_i$, let $\Gamma^k(h) = (S_i^k(h), S_{-i}^k(h))$ be the decision problem that remains after step k of the iterated conditional dominance procedure at h . Suppose that the iterated conditional dominance procedure terminates after K rounds, that is, $\Gamma^{K+1}(h) = \Gamma^K(h)$ for every information set h .

The idea in this proof will be to construct an epistemic model M with the following properties:

- For all players i and every strategy s_i in $S_i^1(\emptyset)$, there is a type $t_i^{s_i}$ for which s_i is optimal (at all information sets $h \in H_i$ that s_i leads to).
- For every $k \geq 1$, if the strategy s_i is in $S_i^k(\emptyset)$, then the associated type $t_i^{s_i}$ expresses up to $(k - 1)$ -fold strong belief in rationality.
- If the strategy s_i is in $S_i^K(\emptyset)$, then the associated type $t_i^{s_i}$ expresses common strong belief in rationality.

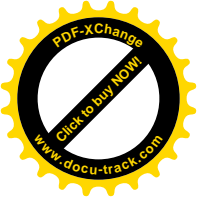
In order to achieve this, we will carry out four steps. In Step 1 we construct, for every strategy s_i in $S_i^1(\emptyset)$, some conditional belief vector about the opponents' strategy choices for which s_i is optimal. In Step 2 we use these conditional belief vectors to construct our epistemic model M . In this model, we define for every strategy s_i in $S_i^1(\emptyset)$ some type $t_i^{s_i}$ for which s_i is optimal. In Step 3 we show that, for every $k \geq 1$ and every s_i in $S_i^k(\emptyset)$, the associated type $t_i^{s_i}$ expresses up to $(k - 1)$ -fold strong belief in rationality. In Step 4 we finally prove that, for every strategy s_i in $S_i^K(\emptyset)$, the associated type $t_i^{s_i}$ expresses common strong belief in rationality.

Step 1. Construction of beliefs.

In this step we construct for every strategy $s_i \in S_i^1(\emptyset)$ some conditional belief vector $b_i^{s_i}$ about the opponents' strategy choices, such that s_i is optimal for $b_i^{s_i}$ at every information set $h \in H_i$ that s_i leads to. For every $k \in \{0, \dots, K - 1\}$, let D_i^k be the set of strategies for player i that are in $S_i^k(\emptyset)$ but not in $S_i^{k+1}(\emptyset)$. To define the conditional belief vectors $b_i^{s_i}$ we distinguish the following two cases.

(1) Consider first some $k \in \{1, \dots, K - 1\}$ and some strategy s_i for player i in D_i^k . By Lemma 9.8.3, we can find at every $h \in H_i$ that s_i leads to, some conditional belief $b_i^{s_i}(h) \in \Delta(S_{-i}^{k-1}(h))$ such that s_i is optimal for $b_i^{s_i}(h)$ (among all strategies in $S_i(h)$).

At every $h \in H_i$ that s_i does not lead to, we construct the conditional belief $b_i^{s_i}(h) \in \Delta(S_{-i}(h))$ as follows: Consider the largest m less than or equal to $k - 1$ such that some opponents' strategy combination in $S_{-i}^m(\emptyset)$ leads to h . Then, choose some arbitrary belief $b_i^{s_i}(h) \in \Delta(S_{-i}^m(\emptyset) \cap S_{-i}(h))$.



In this way, we construct a complete conditional belief vector $b_i^{s_i} = (b_i^{s_i}(h))_{h \in H_i}$ such that s_i is optimal for $b_i^{s_i}(h)$ at all $h \in H_i$ that s_i leads to. Or, in short, s_i is optimal for $b_i^{s_i}$.

(2) Consider next some strategy $s_i \in S_i^K(\emptyset)$ that survives the full iterated conditional dominance procedure. Then, s_i is also in $S_i^{K+1}(\emptyset)$. Hence, by Lemma 9.8.3 we can find at every $h \in H_i$ that s_i leads to, some conditional belief $b_i^{s_i}(h) \in \Delta(S_{-i}^K(h))$ such that s_i is optimal for $b_i^{s_i}(h)$.

At every $h \in H_i$ that s_i does not lead to, we construct the conditional belief $b_i^{s_i}(h) \in \Delta(S_{-i}(h))$ as follows: Consider the largest m less than or equal to K such that some opponents' strategy combination in $S_{-i}^m(\emptyset)$ leads to h . Then, choose some arbitrary belief $b_i^{s_i}(h) \in \Delta(S_{-i}^m(\emptyset) \cap S_{-i}(h))$.

In this way, we construct a complete conditional belief vector $b_i^{s_i} = (b_i^{s_i}(h))_{h \in H_i}$ such that s_i is optimal for $b_i^{s_i}(h)$ at all $h \in H_i$ that s_i leads to.

Step 2. Construction of types.

We will now use these conditional belief vectors $b_i^{s_i}$ to construct, for every strategy $s_i \in S_i$, some type $t_i^{s_i}$. For all players i , let the set of types T_i be given by

$$T_i = \{t_i^{s_i} : s_i \in S_i\}.$$

We will define, for every type $t_i^{s_i}$ and every information set $h \in H_i$, the corresponding conditional belief $b_i(t_i^{s_i}, h)$ on $S_{-i}(h) \times T_{-i}$.

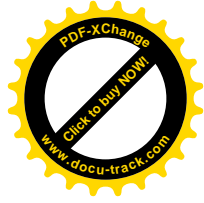
Before we do so, let us denote by T_i^k the set of types $t_i^{s_i}$ with $s_i \in S_i^k(\emptyset)$, for every $k \in \{0, \dots, K\}$. Here, $S_i^k(\emptyset)$ contains those strategies that survive step k of the iterated conditional dominance procedure at \emptyset . By definition, we set $S_i^0(\emptyset) = S_i$. As $S_i^K(\emptyset) \subseteq S_i^{K-1}(\emptyset) \subseteq \dots \subseteq S_i^0(\emptyset)$, it follows that $T_i^K \subseteq T_i^{K-1} \subseteq \dots \subseteq T_i^0$, where $T_i^0 = T_i$.

In order to define the conditional beliefs $b_i(t_i^{s_i}, h)$, we distinguish the following three cases.

(1) Take first a type $t_i^{s_i}$ with $s_i \in D_i^0$. That is, $t_i^{s_i} \in T_i^0 \setminus T_i^1$. We define the conditional beliefs of $t_i^{s_i}$ in an arbitrary way.

(2) Consider next some type $t_i^{s_i}$ with $s_i \in D_i^k$ for some $k \in \{1, \dots, K-1\}$. That is, $t_i^{s_i} \in T_i^k \setminus T_i^{k+1}$ for some $k \in \{1, \dots, K-1\}$. In Step 1 we constructed a conditional belief vector $b_i^{s_i}$ for which s_i is optimal. At every information set $h \in H_i$, let $b_i(t_i^{s_i}, h)$ be the conditional belief about the opponents' strategy-type combinations given by

$$b_i(t_i^{s_i}, h)((s_j, t_j)_{j \neq i}) := \begin{cases} b_i^{s_i}(h)((s_j)_{j \neq i}), & \text{if } t_j = t_j^{s_j} \text{ for every } j \neq i \\ 0, & \text{otherwise.} \end{cases}$$



So, at every $h \in H_i$, type $t_i^{s_i}$ holds the same belief about the opponents' strategy choices as $b_i^{s_i}(h)$. Since strategy s_i is optimal for the conditional belief vector $b_i^{s_i}$, it follows that strategy s_i is optimal for type $t_i^{s_i}$.

Remember that $b_i^{s_i}(h)$ assigns, at every $h \in H_i$ that s_i leads to, only positive probability to opponents' strategy combinations in $S_{-i}^{k-1}(h)$. Hence, at every information set $h \in H_i$ that s_i leads to, type $t_i^{s_i}$ assigns only positive probability to opponents' strategy-type pairs (s_j, t_j) where s_j is in $S_{-i}^{k-1}(h)$ and $t_j = t_j^{s_j}$. By Lemma 9.8.1 we know that at every such information set $h \in H_i$ that s_i leads to, $S_{-i}^{k-1}(h) = S_{-i}^m(\emptyset) \cap S_{-i}(h)$, where m is the highest number less than or equal to $k - 1$ such that $S_{-i}^m(\emptyset) \cap S_{-i}(h)$ is nonempty. As a consequence, at every information set $h \in H_i$ that s_i leads to, type $t_i^{s_i}$ assigns only positive probability to opponents' types $t_j \in T_j^m$, where m is the highest number less than or equal to $k - 1$ such that $S_{-i}^m(\emptyset) \cap S_{-i}(h)$ is nonempty.

(3) Consider finally some type $t_i^{s_i}$ with $s_i \in S_i^K(\emptyset)$. That is, $t_i^{s_i} \in T_i^K$. In Step 1 we constructed a conditional belief vector $b_i^{s_i}$ for which s_i is optimal. At every information set $h \in H_i$, let $b_i(t_i^{s_i}, h)$ be the conditional belief about the opponents' strategy-type combinations given by

$$b_i(t_i^{s_i}, h)((s_j, t_j)_{j \neq i}) := \begin{cases} b_i^{s_i}(h)((s_j)_{j \neq i}), & \text{if } t_j = t_j^{s_j} \text{ for every } j \neq i \\ 0, & \text{otherwise.} \end{cases}$$

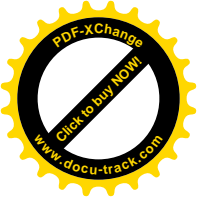
Hence, at every $h \in H_i$, type $t_i^{s_i}$ holds the same belief about the opponents' strategy choices as $b_i^{s_i}(h)$. Since strategy s_i is optimal for the conditional belief vector $b_i^{s_i}$, it follows that strategy s_i is optimal for type $t_i^{s_i}$.

Remember that $b_i^{s_i}(h)$ assigns, at every $h \in H_i$ that s_i leads to, only positive probability to opponents' strategy combinations in $S_{-i}^K(h)$. Hence, at every information set $h \in H_i$ that s_i leads to, type $t_i^{s_i}$ assigns only positive probability to opponents' strategy-type pairs (s_j, t_j) where s_j is in $S_{-i}^K(h)$ and $t_j = t_j^{s_j}$. By Lemma 9.8.1 we know that at every such information set $h \in H_i$ that s_i leads to, $S_{-i}^K(h) = S_{-i}^m(\emptyset) \cap S_{-i}(h)$, where m is the highest number less than or equal to K such that $S_{-i}^m(\emptyset) \cap S_{-i}(h)$ is nonempty. As a consequence, at every information set $h \in H_i$ that s_i leads to, type $t_i^{s_i}$ assigns only positive probability to opponents' types $t_j \in T_j^m$, where m is the highest number less than or equal to K such that $S_{-i}^m(\emptyset) \cap S_{-i}(h)$ is nonempty.

The construction of the epistemic model M is hereby complete.

Step 3. Every type $t_i \in T_i^k$ with $k \geq 1$ expresses up to $(k - 1)$ -fold strong belief rationality.

In order to prove this step, we will show the following lemma.



LEMMA 9.8.4. *For every $k \geq 1$, the following two statements are true:*

- (a) *Every strategy that can rationally be chosen under expressing up to $(k - 1)$ -fold strong belief in rationality must be in $\Gamma^k(\emptyset)$.*
- (b) *For every strategy s_i in $\Gamma^k(\emptyset)$, the associated type $t_i^{s_i}$ in the epistemic model M expresses up to $(k - 1)$ -fold strong belief in rationality.*

Note that, by construction, strategy s_i is optimal for the type $t_i^{s_i}$ if s_i is in $\Gamma^1(\emptyset)$. So, by combining the two statements (a) and (b), we show that for every $k \geq 1$, the decision problem $\Gamma^k(\emptyset)$ contains exactly those strategies that can rationally be chosen under expressing up to $(k-1)$ -fold strong belief in rationality.

Proof of Lemma 9.8.4. We prove the two statements by induction on k .

Induction start. Start with $k = 1$.

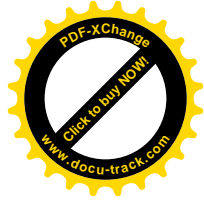
(a) Take a strategy s_i that can rationally be chosen under 0-fold strong belief in rationality. Then, s_i is optimal for some conditional belief vector. So, s_i is not strictly dominated in any full decision problem $\Gamma^0(h)$ at which i is active, and hence s_i is in $\Gamma^1(\emptyset)$.

(b) Take some strategy s_i in $\Gamma^1(\emptyset)$ and consider the associated type $t_i^{s_i}$. By definition, every type expresses 0-fold strong belief in rationality, and so does $t_i^{s_i}$.

Induction step. Take now some $k \in \{2, \dots, K\}$, and assume that the two statements (a) and (b) are true for every $m \leq k - 1$.

(a) Take a strategy s_i that can rationally be chosen under expressing up to $(k - 1)$ -fold strong belief in rationality. So, s_i is optimal for some type t_i that expresses up to $(k - 1)$ -fold strong belief in rationality. Hence, type t_i satisfies the following property at every information set $h \in H_i$: Let m be the highest number, less than or equal to $k - 2$, such that there is an opponents' strategy combination leading to h that can rationally be chosen under expressing up to m -fold strong belief in rationality. Then, $b_i(t_i, h)$ only assigns positive probability to such opponents' strategy combinations.

We know, from the induction assumption, that for every $m \leq k - 2$, the strategies that can rationally be chosen under expressing up to m -fold strong belief in rationality are exactly the strategies in $\Gamma^{m+1}(\emptyset)$. So, type t_i satisfies the following property at every information set $h \in H_i$: If m is the highest number, less than or equal to $k - 1$, such that there is an opponents' strategy combination in $\Gamma^m(\emptyset)$ leading to h , then $b_i(t_i, h)$



only assigns positive probability to opponents' strategy combinations in $\Gamma^m(\emptyset)$ that lead to h .

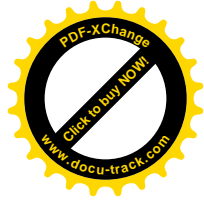
Now, take an information set $h \in H_i$ that strategy s_i leads to. Since strategy s_i can rationally be chosen under expressing up to $(k - 1)$ -fold strong belief in rationality, we have in particular that s_i can rationally be chosen under expressing up to $(k - 2)$ -fold strong belief in rationality. So, by our induction assumption on (a) we know that s_i is in $\Gamma^{k-1}(\emptyset)$. As s_i leads to h , the m we have chosen above is also the highest m less than or equal to $k - 1$ such that there is a players' strategy combination (including player i 's strategy) in $\Gamma^m(\emptyset)$ leading to h . So, type t_i satisfies the following property at every information set $h \in H_i$ that strategy s_i leads to: If m is the highest number, less than or equal to $k - 1$, such that there is a players' strategy combination in $\Gamma^m(\emptyset)$ leading to h , then $b_i(t_i, h)$ only assigns positive probability to opponents' strategy combinations in $\Gamma^m(\emptyset)$ that lead to h . However, by Lemma 9.8.1, the opponents' strategy combinations in $\Gamma^m(\emptyset)$ that lead to h are exactly the opponents' strategy combinations in $\Gamma^{k-1}(h)$.

Hence, we conclude that type t_i assigns, at every information set $h \in H_i$ that s_i leads to, only positive probability to opponents' strategy combinations in $\Gamma^{k-1}(h)$. Since strategy s_i is optimal for type t_i at every $h \in H_i$ that s_i leads to, we may conclude that s_i is not strictly dominated within any decision problem $\Gamma^{k-1}(h)$ at which i is active. But then, by definition of the iterated conditional dominance procedure, we conclude that s_i is in $\Gamma^k(\emptyset)$, which was to show.

(b) Take now some strategy s_i in $\Gamma^k(\emptyset)$, and consider the associated type $t_i^{s_i}$. We will prove that $t_i^{s_i}$ expresses up to $(k - 1)$ -fold strong belief in rationality. By our induction assumption, $t_i^{s_i}$ expresses up to $(k - 2)$ -fold strong belief in rationality. So, it remains to show that $t_i^{s_i}$ expresses $(k - 1)$ -fold strong belief in rationality. Consider an information set $h \in H_i$ and suppose that there is a combination of opponents' types, possibly outside M , that express up to $(k - 2)$ -fold strong belief in rationality, and for which there is a combination of optimal strategies leading to h . Then, we have to verify two conditions:

(1) The epistemic model M must contain at least one such combination of opponents' types.

(2) Type $t_i^{s_i}$ must at h only assign positive probability to opponents' strategy-type combinations where the strategy combinations lead to h , the types express up to $(k - 2)$ -fold strong belief in rationality, and the strategies are optimal for the types.



We first verify condition (1). So, we assume that for every opponent j there is some type t_j (possibly outside M) expressing up to $(k - 2)$ -fold strong belief in rationality, and some strategy s_j which is optimal for that type, such that this combination of opponents' strategies leads to h . By our induction assumption on (a), it follows that s_j must be in $\Gamma^{k-1}(\emptyset)$. But then, by our induction assumption on (b), it follows that the associated type $t_j^{s_j}$ in M expresses up to $(k - 2)$ -fold strong belief in rationality. Moreover, we know that s_j is optimal for $t_j^{s_j}$. Hence, M contains a combination of opponents' types that express up to $(k - 2)$ -fold strong belief in rationality, and for which there is a combination of optimal strategies leading to h , namely the combination of types $t_j^{s_j}$ for every opponent j constructed above. Hence, condition (1) is satisfied.

We now verify condition (2). Again, assume that at h there is a combination of opponents' types that express up to $(k - 2)$ -fold strong belief in rationality, and for which there is a combination of optimal strategies leading to h . By our induction assumption on (a), we know that such opponents' strategies must be in $\Gamma^{k-1}(\emptyset)$. Hence, there is a combination of opponents' strategies in $\Gamma^{k-1}(\emptyset)$ that leads to h .

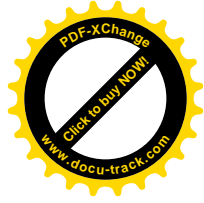
We show that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type combinations where the strategy combinations lead to h , the types express up to $(k - 2)$ -fold strong belief in rationality, and the strategies are optimal for the types. We distinguish two cases.

Case 1. Suppose that strategy s_i leads to h .

Since, by assumption, s_i is in $\Gamma^k(\emptyset)$, it must be that $s_i \in D_i^{k'}$ for some $k' \geq k$. But then, by construction of the types, $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $\Gamma^{k'-1}(h)$, since s_i leads to h . As $\Gamma^{k'-1}(h)$ is contained in $\Gamma^{k-1}(h)$, we conclude that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $\Gamma^{k-1}(h)$.

Above we have seen that there is a combination of opponents' strategies in $\Gamma^{k-1}(\emptyset)$ that leads to h . Moreover, as s_i is in $\Gamma^k(\emptyset)$ and s_i leads to h , we know that there is a player i strategy in $\Gamma^{k-1}(\emptyset)$ that leads to h , namely s_i . So, in total, we conclude that there is a players' strategy combination in $\Gamma^{k-1}(\emptyset)$ that leads to h . Hence, we know by Lemma 9.8.1 that $\Gamma^{k-1}(h)$ contains precisely those opponents' strategy combinations in $\Gamma^{k-1}(\emptyset)$ that lead to h . So, type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategies in $\Gamma^{k-1}(\emptyset)$.

Moreover, by construction of the types, $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs (s_j, t_j) with $t_j = t_j^{s_j}$. Together with the above, we may thus conclude that type $t_i^{s_i}$ assigns at h



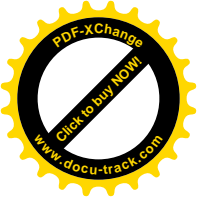
only positive probability to opponents' strategy-type pairs (s_j, t_j) where s_j is in $\Gamma^{k-1}(\emptyset)$ and $t_j = t_j^{s_j}$. By our induction assumption on (b), we know that every such type $t_j^{s_j}$ expresses up to $(k-2)$ -fold strong belief in rationality. We also know that for every s_j in $\Gamma^{k-1}(\emptyset)$, the strategy s_j is optimal for the type $t_j^{s_j}$ since $s_j \in D_j^m$ for some $m \geq 1$. Putting all these facts together, it follows that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs $(s_j, t_j^{s_j})$ where s_j is optimal for $t_j^{s_j}$ and type $t_j^{s_j}$ expresses up to $(k-2)$ -fold strong belief in rationality. This is what we had to show.

Case 2. Suppose that strategy s_i does not lead to h .

Since, by assumption, s_i is in $\Gamma^k(\emptyset)$, it must be that $s_i \in D_i^{k'}$ for some $k' \geq k$. As s_i does not lead to h we have, by our construction in Step 1, that the conditional belief $b_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $S_{-i}^m(\emptyset)$, where m is the largest number less than or equal to $k' - 1$ such that some opponents' strategy combination in $S_{-i}^m(\emptyset)$ leads to h . Remember our assumption above that there is a combination of opponents' strategies in $\Gamma^{k-1}(\emptyset)$ that leads to h . In other words, there is an opponents' strategy combination in $S_{-i}^{k-1}(\emptyset)$ leading to h , which means that $m \geq k - 1$. Since $S_{-i}^m(\emptyset) \subseteq S_{-i}^{k-1}(\emptyset)$, we conclude that the conditional belief $b_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $S_{-i}^{k-1}(\emptyset)$. By construction, the type $t_i^{s_i}$ holds at h the same belief about the opponents' strategies as $b_i^{s_i}$, and hence also type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $S_{-i}^{k-1}(\emptyset)$.

Moreover, the type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs (s_j, t_j) with $t_j = t_j^{s_j}$. Together with the above, we may thus conclude that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs (s_j, t_j) where s_j is in $S_j^{k-1}(\emptyset)$ and $t_j = t_j^{s_j}$. By our induction assumption on (b), we know that every such type $t_j^{s_j}$ expresses up to $(k-2)$ -fold strong belief in rationality. We also know that for every s_j in $S_j^{k-1}(\emptyset)$, the strategy s_j is optimal for the type $t_j^{s_j}$ since $s_j \in D_j^m$ for some $m \geq 1$. Putting all these facts together, it follows that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs $(s_j, t_j^{s_j})$ where s_j is optimal for $t_j^{s_j}$ and type $t_j^{s_j}$ expresses up to $(k-2)$ -fold strong belief in rationality. This is what we had to show.

Overall, we see that type $t_i^{s_i}$ satisfies conditions (1) and (2) above, and hence expresses $(k-1)$ -fold strong belief rationality. Since we already knew, by our induction assumption, that $t_i^{s_i}$ expresses up to $(k-2)$ -fold



strong belief in rationality, it follows that $t_i^{s_i}$ expresses up to $(k - 1)$ -fold strong belief in rationality. So, for every strategy s_i in $\Gamma^k(\emptyset)$, the associated type $t_i^{s_i}$ expresses up to $(k - 1)$ -fold strong belief rationality. This proves (b).

By induction on k , the statements (a) and (b) hold for every $k \geq 1$. This completes the proof Lemma 9.8.4. \diamond

With Lemma 9.8.4 we can now easily show the statement in step 3. Take, namely, some type $t_i \in T_i^k$. Then, $t_i = t_i^{s_i}$ for some s_i in $\Gamma^k(\emptyset)$. By Lemma 9.8.4, part (b), it follows that t_i expresses up to $(k - 1)$ -fold strong belief in rationality. So, Step 3 is complete now.

Step 4. Every type $t_i \in T_i^K$ expresses common strong belief in rationality.

In order to prove this, we show the following lemma.

LEMMA 9.8.5. *For every $k \geq K - 1$, every type $t_i \in T_i^K$ expresses up to k -fold strong belief in rationality.*

Proof of Lemma 9.8.5. We prove the statement by induction on k .

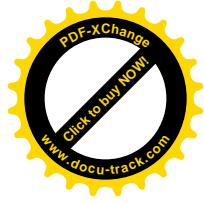
Induction start. Start with $k = K - 1$. Take some $t_i \in T_i^K$. Then, $t_i = t_i^{s_i}$ for some s_i in $\Gamma^K(\emptyset)$. By Lemma 9.8.4 we know that type $t_i^{s_i}$ expresses up to $(K - 1)$ -fold strong belief in rationality.

Induction step. Take now some $k \geq K$, and assume that for all players i , every type $t_i \in T_i^K$ expresses up to $(k - 1)$ -fold strong belief in rationality.

Consider an arbitrary type $t_i \in T_i^K$. That is, $t_i = t_i^{s_i}$ for some s_i in $\Gamma^K(\emptyset)$. We must show that $t_i^{s_i}$ expresses up to k -fold strong belief in rationality. Since, by our induction assumption, $t_i^{s_i}$ expresses up to $(k - 1)$ -fold strong belief in rationality, it is sufficient to show that $t_i^{s_i}$ expresses k -fold strong belief in rationality. Consider an information set $h \in H_i$ and suppose that there is a combination of opponents' types, possibly outside M , that express up to $(k - 1)$ -fold strong belief in rationality, and for which there is a combination of optimal strategies leading to h . Then, we have to verify two conditions:

(1) The epistemic model M must contain at least one such combination of opponents' types.

(2) Type $t_i^{s_i}$ must at h only assign positive probability to opponents' strategy-type combinations where the strategy combinations lead to h , the types express up to $(k - 1)$ -fold strong belief in rationality, and the strategies are optimal for the types.



We first verify condition (1). So, we assume that for every opponent j there is some type t_j (possibly outside M) expressing up to $(k - 1)$ -fold strong belief in rationality, and some strategy s_j which is optimal for that type, such that this combination of opponents' strategies leads to h . By Lemma 9.8.4 we know that every such strategy s_j must be in $\Gamma^K(\emptyset)$, as $k \geq K$. But then, by our induction assumption, it follows that the associated type $t_j^{s_j}$ in M expresses up to $(k - 1)$ -fold strong belief in rationality. Moreover, we know that s_j is optimal for $t_j^{s_j}$. Hence, M contains a combination of opponents' types that express up to $(k - 1)$ -fold strong belief in rationality, and for which there is a combination of optimal strategies leading to h , namely the combination of types $t_j^{s_j}$ for every opponent j . Hence, condition (1) is satisfied.

We now verify condition (2). Assume that at h there is a combination of opponents' types that express up to $(k - 1)$ -fold strong belief in rationality, and for which there is a combination of optimal strategies leading to h . By Lemma 9.8.4, we know that such opponents' strategies must be in $\Gamma^K(\emptyset)$. Hence, there is a combination of opponents' strategies in $\Gamma^K(\emptyset)$ that leads to h .

We show that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type combinations where the strategy combinations lead to h , the types express up to $(k - 1)$ -fold strong belief in rationality, and the strategies are optimal for the types. Again, we distinguish two cases.

Case 1. Suppose that strategy s_i leads to h .

Since, by assumption, s_i is in $\Gamma^K(\emptyset)$, we have that $s_i \in S_i^K(\emptyset)$. As s_i leads to h we have, by construction, that $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $\Gamma^K(h)$. Above we have seen that there is a combination of opponents' strategies in $\Gamma^K(\emptyset)$ that leads to h . Moreover, as s_i is in $\Gamma^K(\emptyset)$ and s_i leads to h , we know that there is a player i strategy in $\Gamma^K(\emptyset)$ that leads to h , namely s_i . So, in total, we conclude that there is a players' strategy combination in $\Gamma^K(\emptyset)$ that leads to h . Hence, we know by Lemma 9.8.1 that $\Gamma^K(h)$ contains precisely those opponents' strategy combinations in $\Gamma^K(\emptyset)$ that lead to h . So, type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategies s_j in $\Gamma^K(\emptyset)$.

Moreover, by construction of the types, $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs (s_j, t_j) with $t_j = t_j^{s_j}$. Together with the above, we may thus conclude that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs (s_j, t_j) where s_j is in $\Gamma^K(\emptyset)$ and $t_j = t_j^{s_j}$. By our induction assumption, we know that every such type $t_j^{s_j}$ expresses up to $(k - 1)$ -fold strong belief in rationality.



We also know that for every s_j in $\Gamma^K(\emptyset)$, the strategy s_j is optimal for the type $t_j^{s_j}$. Putting all these facts together, it follows that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs $(s_j, t_j^{s_j})$ where s_j is optimal for $t_j^{s_j}$ and type $t_j^{s_j}$ expresses up to $(k-1)$ -fold strong belief in rationality. This is what we had to show.

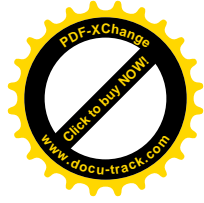
Case 2. Suppose that strategy s_i does not lead to h .

Since, by assumption, s_i is in $\Gamma^K(\emptyset)$, we have that $s_i \in S_i^K(\emptyset)$. As s_i does not lead to h we have, by our construction in Step 1, that the conditional belief $b_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $S_{-i}^m(\emptyset)$, where m is the largest number less than or equal to K such that some opponents' strategy combination in $S_{-i}^m(\emptyset)$ leads to h . Remember our assumption above that there is a combination of opponents' strategies in $\Gamma^K(\emptyset)$ that leads to h . In other words, there is an opponents' strategy combination in $S_{-i}^K(\emptyset)$ leading to h , which means that $m = K$. Hence, we conclude that the conditional belief $b_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $S_{-i}^K(\emptyset)$. By construction, the type $t_i^{s_i}$ holds at h the same belief about the opponents' strategies as $b_i^{s_i}$, and hence also type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy combinations in $S_{-i}^K(\emptyset)$.

Moreover, the type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs (s_j, t_j) with $t_j = t_j^{s_j}$. Together with the above, we may thus conclude that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs (s_j, t_j) where s_j is in $S_j^K(\emptyset)$ and $t_j = t_j^{s_j}$. By our induction assumption, we know that every such type $t_j^{s_j}$ expresses up to $(k-1)$ -fold strong belief in rationality. We also know that for every s_j in $S_j^K(\emptyset)$, the strategy s_j is optimal for the type $t_j^{s_j}$. Putting all these facts together, it follows that type $t_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs $(s_j, t_j^{s_j})$ where s_j is optimal for $t_j^{s_j}$ and type $t_j^{s_j}$ expresses up to $(k-1)$ -fold strong belief in rationality. This is what we had to show.

Overall, we see that type $t_i^{s_i}$ satisfies conditions (1) and (2) above, and hence expresses k -fold strong belief rationality. Since we already knew, by our induction assumption, that $t_i^{s_i}$ expresses up to $(k-1)$ -fold strong belief in rationality, it follows that $t_i^{s_i}$ expresses up to k -fold strong belief in rationality. So, for every strategy s_i in $\Gamma^K(\emptyset)$, the associated type $t_i^{s_i}$ expresses up to k -fold strong belief rationality. By induction on k , the statement in the lemma follows. \diamond

Obviously, the statement in Step 4 follows from Lemma 9.8.5, and hence Step 4 is hereby complete.



With Lemma 9.8.4 and Lemma 9.8.5 at hand it is now easy to prove Theorem 9.3.3. Take, namely, some strategy s_i that can rationally be chosen under expressing up to k -fold strong belief in rationality, for some $k \geq 0$. Then, by Lemma 9.8.4, part (a), it follows that s_i is in $\Gamma^{k+1}(\emptyset)$, so s_i survives $k + 1$ rounds of the iterated conditional dominance procedure at \emptyset . On the other hand, take some strategy in $\Gamma^{k+1}(\emptyset)$. Then, by Lemma 9.8.4, part (b), we know that the associated type $t_i^{s_i}$ expresses up to k -fold strong belief in rationality. As s_i is optimal for $t_i^{s_i}$, we conclude that s_i can rationally be chosen under expressing up to k -fold strong belief in rationality. So, we see that $\Gamma^{k+1}(\emptyset)$ contains precisely those strategies that can rationally be chosen under expressing up to k -fold strong belief in rationality. This proves part (1) of Theorem 9.3.3.

Now, take some strategy s_i that can rationally be chosen under common strong belief in rationality. Then, in particular, s_i can rationally be chosen under expressing up to $(K - 1)$ -fold strong belief in rationality. Hence, by Lemma 9.8.4, part (a), it follows that s_i is in $\Gamma^K(\emptyset)$, so s_i survives the iterated conditional dominance procedure at \emptyset . So, every strategy s_i that can rationally be chosen under common strong belief in rationality survives the iterated conditional dominance procedure at \emptyset .

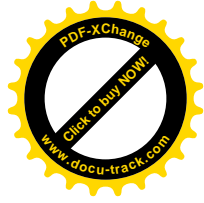
On the other hand, take some strategy s_i that survives the iterated conditional dominance procedure at \emptyset . Hence, s_i is in $\Gamma^K(\emptyset)$. By Lemma 9.8.5 we know that the associated type $t_i^{s_i}$ expresses k -fold strong belief in rationality for all k . In other words, the associated type $t_i^{s_i}$ expresses common strong belief in rationality. As s_i is optimal for $t_i^{s_i}$, it follows that s_i can rationally be chosen under common strong belief in rationality. So, every strategy s_i that survives the iterated conditional dominance procedure at \emptyset can rationally be chosen under common strong belief in rationality.

Together with our conclusion above, we see that a strategy s_i can rationally be chosen under common strong belief in rationality, if and only if, s_i survives the iterated conditional dominance procedure at \emptyset . This proves part (2) of Theorem 9.3.3. The proof of Theorem 9.3.3 is hereby complete. ■

Outcomes under Common Strong Belief in Rationality and Common Belief in Future Rationality. We will finally prove the following theorem.

THEOREM 9.4.2 (*Outcomes under common strong belief in rationality and common belief in future rationality*)

Every outcome you initially deem possible under common strong belief



in rationality, is also initially deemed possible under common belief in future rationality.

Proof. We will use the following terminology and notation in the proof. For every information set h , let $\Gamma^{icd}(h)$ be the reduced game that remains at h after applying the iterated conditional dominance procedure, and let $\Gamma^{bd}(h)$ be the reduced game that remains at h after applying the backward dominance procedure. We say that an information set h is *reachable* under common strong belief in rationality if there is a strategy combination in $\Gamma^{icd}(\emptyset)$ that leads to h . This makes sense as we know from Theorem 9.3.3 that $\Gamma^{icd}(\emptyset)$ contains precisely those strategies that can rationally be chosen under common strong belief in rationality. Similarly, an information set h is called *reachable* under common belief in future rationality if there is a strategy combination in $\Gamma^{bd}(\emptyset)$ that leads to h . Denote by H^{icd} the collection of information sets that are reachable under common strong belief in rationality, and let H_i^{bd} be the collection of information sets for player i that are reachable under common belief in future rationality. Finally, an outcome z is said to be reachable under common strong belief in rationality if there is a strategy combination in $\Gamma^{icd}(\emptyset)$ leading to z . Similarly, an outcome z is called reachable under common belief in future rationality if there is a strategy combination in $\Gamma^{bd}(\emptyset)$ leading to z . Hence, we must prove that every outcome z that is reachable under common strong belief in rationality is also reachable under common belief in future rationality.

We proceed by the following steps. In Step 1 we transform every strategy s_i in $\Gamma^{icd}(\emptyset)$ into some new strategy $\sigma_i(s_i)$, and show that it is optimal for some particular belief vector $b_i(\sigma_i(s_i))$. In Step 2, we use the transformed strategies and belief vectors to construct, for every s_i in $\Gamma^{icd}(\emptyset)$, some type $t_i^{\sigma_i(s_i)}$. In Step 3 we prove that, for every s_i in $\Gamma^{icd}(\emptyset)$, strategy $\sigma_i(s_i)$ is optimal for type $t_i^{\sigma_i(s_i)}$. In Step 4 we prove that every type $t_i^{\sigma_i(s_i)}$ so constructed expresses common belief in future rationality. Hence, every strategy $\sigma_i(s_i)$ induced by some strategy s_i in $\Gamma^{icd}(\emptyset)$ can rationally be chosen under common belief in future rationality, and is therefore in $\Gamma^{bd}(\emptyset)$. In Step 5 we finally show that, whenever a strategy combination (s_1, \dots, s_n) in $\Gamma^{icd}(\emptyset)$ leads to an outcome z , then the induced strategy combination $(\sigma_1(s_1), \dots, \sigma_n(s_n))$ in $\Gamma^{bd}(\emptyset)$ leads to outcome z as well. That is, every outcome z that is reachable under common strong belief in rationality is also reachable under common belief in future rationality.

Step 1. Transformation of strategies.



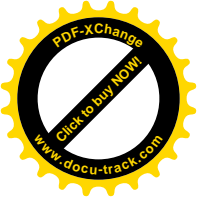
Consider a strategy s_i in $\Gamma^{icd}(\emptyset)$. Remember that H_i^{icd} contains those information sets $h \in H_i$ that are reachable under common strong belief in rationality. That is, H_i^{icd} contains those information sets $h \in H_i$ for which there is some strategy combination in $\Gamma^{icd}(\emptyset)$ leading to h . Then, we know from the optimality principle in Lemma 9.8.3 that, for every $h \in H_i^{icd}$ that s_i leads to, there is some belief $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{icd}(h))$ such that s_i is optimal at h for $b_i(\sigma_i(s_i), h)$. Here, $S_{-i}^{icd}(h)$ denotes the set of opponents' strategy combinations in $\Gamma^{icd}(h)$. Now, take an information set $h \in H_i^{icd}$ that s_i leads to. Then, by definition, there is a strategy combination in $\Gamma^{icd}(\emptyset)$ that leads to h . Hence, by Lemma 9.8.1 we know that $S_{-i}^{icd}(h) = S_{-i}^{icd}(\emptyset) \cap S_{-i}(h)$. So, for every $h \in H_i^{icd}$ that s_i leads to, there is a belief $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{icd}(\emptyset) \cap S_{-i}(h))$, such that s_i is optimal at h for $b_i(\sigma_i(s_i), h)$.

Consider now an information set $h \in H_i$ that s_i leads to, such that (a) $h \notin H_i^{icd}$, and (b) every $h' \in H_i$ preceding h , if there is any, is in H_i^{icd} . So, h is a first information set for player i that is not reachable under common strong belief in rationality. Then, it is clear that every information set that follows h is also not reachable under common strong belief in rationality. Now, define at h an arbitrary belief $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{bd}(h))$, where $S_{-i}^{bd}(h)$ is the set of opponents' strategy combinations in $\Gamma^{bd}(h)$. By Lemma 8.14.5 in the proofs section of Chapter 8, we can find for every $h' \in H_i$ following h a belief $b_i(\sigma_i(s_i), h') \in \Delta(S_{-i}^{bd}(h'))$, and a strategy $\tilde{s}_i(h)$ leading to h , such that $\tilde{s}_i(h)$ is optimal at every $h' \in H_i$ weakly following h for the belief $b_i(\sigma_i(s_i), h')$.

Now, let $\sigma_i(s_i)$ be the strategy that (a) at every $h \in H_i^{icd}$ that $\sigma_i(s_i)$ leads to, selects the same choice as s_i , and (b) at every $h \notin H_i^{icd}$ that $\sigma_i(s_i)$ leads to, selects the same choice as $\tilde{s}_i(h')$, where h' is the first information set for player i weakly preceding h which is not in H_i^{icd} .

Above, we have constructed for every $h \in H_i$ that $\sigma_i(s_i)$ leads to, some belief $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}(h))$. We will show that, in fact, $\sigma_i(s_i)$ is optimal at each of these information sets h for the belief $b_i(\sigma_i(s_i), h)$. We distinguish two cases.

Case 1. Suppose that $h \in H_i^{icd}$. Then we have, by construction, that $\sigma_i(s_i)$ coincides with s_i on h , and on every $h' \in H_i^{icd}$ that follows h . Moreover, $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{icd}(\emptyset) \cap S_{-i}(h))$ as we have seen. But then, every information set $h' \in H_i$ that can be reached with positive probability under $\sigma_i(s_i)$ and $b_i(\sigma_i(s_i), h)$ is in H_i^{icd} . So, to evaluate the optimality of $\sigma_i(s_i)$ at h under the belief $b_i(\sigma_i(s_i), h)$, we only need the choices that $\sigma_i(s_i)$ prescribes at information sets in H_i^{icd} . But these choices are the same as the choices prescribed by s_i . Since we know that s_i is optimal at



h for the belief $b_i(\sigma_i(s_i), h)$, we conclude that also $\sigma_i(s_i)$ is optimal at h for the belief $b_i(\sigma_i(s_i), h)$.

Case 2. Suppose that $h \notin H_i^{icd}$. Let h' be the first information set for player i weakly preceding h which is not in H_i^{icd} . Then, by construction, $\sigma_i(s_i)$ coincides with $\tilde{s}_i(h')$ at h , and at all player i information sets that follow h . Since we know that $\tilde{s}_i(h')$ is optimal at h for the belief $b_i(\sigma_i(s_i), h)$, it follows that also $\sigma_i(s_i)$ is optimal at h for the belief $b_i(\sigma_i(s_i), h)$.

Finally, define at every $h \in H_i$ that $\sigma_i(s_i)$ does not lead to, some arbitrary belief $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{bd}(h))$. Then, we obtain a complete conditional belief vector $b_i(\sigma_i(s_i))$ for player i .

Summarizing, we have transformed the strategy $s_i \in S_i^{icd}(\emptyset)$ into a new strategy $\sigma_i(s_i)$, and we have defined a conditional belief vector $b_i(\sigma_i(s_i))$, such that

- the new strategy $\sigma_i(s_i)$ coincides with s_i at all information sets in H_i^{icd} that $\sigma_i(s_i)$ leads to,
- for every $h \in H_i^{icd}$ that $\sigma_i(s_i)$ leads to, we have that $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{icd}(\emptyset) \cap S_{-i}(h))$,
- for every other information set $h \in H_i$, we have that $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{bd}(h))$,
- the new strategy $\sigma_i(s_i)$ is optimal for the belief vector $b_i(\sigma_i(s_i))$.

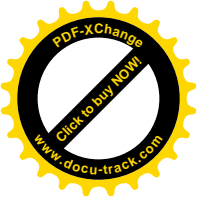
Step 2. Construction of types.

For every player i we will construct a set of types T_i which consists of (a) types $t_i^{\sigma_i(s_i)}$ for every strategy $s_i \in S_i^{icd}(\emptyset)$, and (b) types $\tau_i^{s_i}$ for every strategy $s_i \in S_i$.

(a) We start by defining the beliefs for the types $t_i^{\sigma_i(s_i)}$ for every strategy $s_i \in S_i^{icd}(\emptyset)$. Take some strategy $s_i \in S_i^{icd}(\emptyset)$, and let $\sigma_i(s_i)$ be the transformed strategy constructed in Step 1.

From above, we know that at every $h \in H_i^{icd}$ that $\sigma_i(s_i)$ leads to, strategy $\sigma_i(s_i)$ is optimal for some belief $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{icd}(\emptyset) \cap S_{-i}(h))$. So, the belief $b_i(\sigma_i(s_i), h)$ only assigns positive probability to opponents' strategies s'_j in $S_j^{icd}(\emptyset)$. At every $h \in H_i^{icd}$ that $\sigma_i(s_i)$ leads to, define the conditional belief $b_i(t_i^{\sigma_i(s_i)}, h)$ about the opponents' strategy-type pairs by

$$b_i(t_i^{\sigma_i(s_i)}, h)((s'_j, t_j)_{j \neq i}) := \begin{cases} b_i(\sigma_i(s_i), h)((s_j)_{j \neq i}), & \text{if } s'_j = \sigma_j(s_j) \text{ and} \\ & t_j = t_j^{\sigma_j(s_j)} \text{ for} \\ & \text{every } j \neq i \\ 0, & \text{otherwise.} \end{cases}$$



Hence, at every $h \in H_i^{icd}$ that $\sigma_i(s_i)$ leads to, type $t_i^{\sigma_i(s_i)}$ only assigns positive probability to strategy-type pairs $(\sigma_j(s_j), t_j^{\sigma_j(s_j)})$ with $s_j \in S_j^{icd}(\emptyset)$.

At every other information set $h \in H_i$, we know from above that strategy $\sigma_i(s_i)$ is optimal for some belief $b_i(\sigma_i(s_i), h) \in \Delta(S_{-i}^{bd}(h))$, if $\sigma_i(s_i)$ leads to h . At each of these other information sets $h \in H_i$, define the conditional belief $b_i(t_i^{\sigma_i(s_i)}, h)$ about the opponents' strategy-type pairs by

$$b_i(t_i^{\sigma_i(s_i)}, h)((s_j, t_j)_{j \neq i}) := \begin{cases} b_i(\sigma_i(s_i), h)((s_j)_{j \neq i}), & \text{if } t_j = \tau_j^{s_j} \\ & \text{for every } j \neq i \\ 0, & \text{otherwise.} \end{cases}$$

Hence, at each of these other information sets $h \in H_i$, type $t_i^{\sigma_i(s_i)}$ holds the same belief about the opponents' strategy choices as $b_i(\sigma_i(s_i), h)$, and only assigns positive probability to strategy-type pairs $(s_j, \tau_j^{s_j})$ with $s_j \in S_j^{bd}(h)$.

(b) We now define the beliefs for the types $\tau_i^{s_i}$ for every strategy $s_i \in S_i$. For every strategy s_i , let $H_i^{bd}(s_i)$ be the (possibly empty) collection of information sets $h \in H_i$ for which s_i is in $\Gamma^{bd}(h)$. That is, $H_i^{bd}(s_i)$ contains all those information sets for player i at which s_i survives the backward dominance procedure. By Lemma 8.14.6 in the proofs section of Chapter 8, we can find for every $s_i \in S_i$ some conditional belief vector $\beta_i(s_i) = (\beta_i(s_i, h))_{h \in H_i}$ such that (a) $\beta_i(s_i, h) \in \Delta(S_{-i}^{bd}(h))$ for every $h \in H_i$, and (b) s_i is optimal at every $h \in H_i^{bd}(s_i)$ for the belief $\beta_i(s_i, h)$.

Now, take a given $s_i \in S_i$ and an information set $h \in H_i$. For type $\tau_i^{s_i}$, let $b_i(\tau_i^{s_i}, h)$ be the conditional belief about the opponents' strategy-type pairs given by

$$b_i(\tau_i^{s_i}, h)((s_j, t_j)_{j \neq i}) := \begin{cases} \beta_i(s_i, h)((s_j)_{j \neq i}), & \text{if } t_j = \tau_j^{s_j} \text{ for every } j \neq i \\ 0, & \text{otherwise.} \end{cases}$$

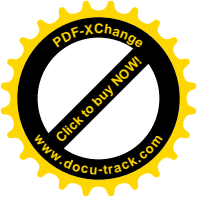
Hence, at every information set $h \in H_i$, type $\tau_i^{s_i}$ holds the same belief about the opponents' strategy choices as $\beta_i(s_i, h)$, and only assigns positive probability to strategy-type pairs $(s_j, \tau_j^{s_j})$ with $s_j \in S_j^{bd}(h)$.

This completes the description of the epistemic model.

Step 3. Strategy $\sigma_i(s_i)$ is optimal for type $t_i^{\sigma_i(s_i)}$.

Take some strategy $s_i \in S_i^{icd}(\emptyset)$. We show that the induced strategy $\sigma_i(s_i)$ is optimal for type $t_i^{\sigma_i(s_i)}$ at every $h \in H_i$ that $\sigma_i(s_i)$ leads to.

Consider first an information set $h \in H_i^{icd}$. Then, by construction, the belief $b_i(t_i^{\sigma_i(s_i)}, h)$ only assigns positive probability to opponents' strategy combinations $(\sigma_j(s_j))_{j \neq i}$ with $s_j \in S_j^{icd}(\emptyset)$ for all j . In Step 1 we have seen that these strategies $\sigma_j(s_j)$ coincide with s_j on information sets in H_j^{icd} , and that strategy $\sigma_i(s_i)$ coincides with s_i on information sets in



H_i^{icd} . So, strategy $\sigma_i(s_i)$, in combination with the belief $b_i(t_i^{\sigma_i(s_i)}, h)$, can only lead with positive probability to information sets in H_i^{icd} .

By construction, the probability that the belief $b_i(t_i^{\sigma_i(s_i)}, h)$ assigns to an opponents' strategy combination $(\sigma_j(s_j))_{j \neq i}$ is the same as the probability that the belief $b_i(\sigma_i(s_i), h)$ assigns to $(s_j)_{j \neq i}$. As $\sigma_j(s_j)$ coincides with s_j on H_j^{icd} , and since we know from Step 1 that strategy $\sigma_i(s_i)$ is optimal for the belief $b_i(\sigma_i(s_i), h)$ at h , it follows that strategy $\sigma_i(s_i)$ is also optimal for the belief $b_i(t_i^{\sigma_i(s_i)}, h)$ at h .

Consider now an information set $h \in H_i$ which is not in H_i^{icd} . Then, by construction, $b_i(t_i^{\sigma_i(s_i)}, h)$ holds the same belief about the opponents' strategy choices as $b_i(\sigma_i(s_i), h)$. From Step 1 we know that strategy $\sigma_i(s_i)$ is optimal for $b_i(\sigma_i(s_i), h)$ at h . Hence, strategy $\sigma_i(s_i)$ is then also optimal for $b_i(t_i^{\sigma_i(s_i)}, h)$ at h .

Summarizing, we may conclude that strategy $\sigma_i(s_i)$ is optimal for the belief $b_i(t_i^{\sigma_i(s_i)}, h)$ at every $h \in H_i$ that $\sigma_i(s_i)$ leads to. That is, $\sigma_i(s_i)$ is optimal for type $t_i^{\sigma_i(s_i)}$.

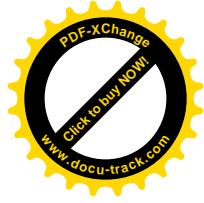
Step 4. All types express common belief in future rationality.

We next show that every type constructed above expresses common belief in future rationality. To do so, it is sufficient to prove that every type in the epistemic model believes in the opponents' future rationality. We distinguish two cases.

Case 1. Consider first a type $t_i^{\sigma_i(s_i)}$ for some $s_i \in S_i^{icd}(\emptyset)$. We show that $t_i^{\sigma_i(s_i)}$ believes in the opponents' future rationality.

Take first some information set $h \in H_i^{icd}$ that $\sigma_i(s_i)$ leads to. We have seen that at h , type $t_i^{\sigma_i(s_i)}$ only assigns positive probability to strategy-type pairs $(\sigma_j(s_j), t_j^{\sigma_j(s_j)})$ with $s_j \in S_j^{icd}(\emptyset)$. From Step 3 we know that strategy $\sigma_j(s_j)$ is optimal for $t_j^{\sigma_j(s_j)}$ at all $h' \in H_j$ that $\sigma_j(s_j)$ leads to. In particular, $\sigma_j(s_j)$ is optimal for $t_j^{\sigma_j(s_j)}$ at all $h' \in H_j$ weakly following h that $\sigma_j(s_j)$ leads to. Hence, at information set h , type $t_i^{\sigma_i(s_i)}$ only assigns positive probability to strategy-type pairs $(\sigma_j(s_j), t_j^{\sigma_j(s_j)})$ where $\sigma_j(s_j)$ is optimal for $t_j^{\sigma_j(s_j)}$ at all $h' \in H_j$ weakly following h that $\sigma_j(s_j)$ leads to. That is, type $t_i^{\sigma_i(s_i)}$ believes at h in the opponents' future rationality.

Take now some other information set $h \in H_i$. By construction, the belief $b_i(t_i^{\sigma_i(s_i)}, h)$ only assigns positive probability to opponents' strategy-type pairs $(s_j, \tau_j^{s_j})$ with $s_j \in S_j^{bd}(h)$.



Take some $s_j \in S_j^{bd}(h)$. Then, by construction of the backward dominance procedure, we have that $s_j \in S_j^{bd}(h')$ for every h' weakly following h that s_j leads to. In other words, if $s_j \in S_j^{bd}(h)$, then every $h' \in H_j$ weakly following h that s_j leads to is in $H_j^{bd}(s_j)$. Remember from above that $H_j^{bd}(s_j)$ is the collection of information sets $h' \in H_j$ with $s_j \in S_j^{bd}(h')$. By construction, at every $h' \in H_j^{bd}(s_j)$, type $\tau_j^{s_j}$ holds the same belief about the opponents' strategy choices as $\beta_j(s_j, h')$. Moreover, at every $h' \in H_j^{bd}(s_j)$, strategy s_j is optimal under the belief $\beta_j(s_j, h')$. So, at every $h' \in H_j^{bd}(s_j)$, strategy s_j is optimal for type $\tau_j^{s_j}$. Since we have seen that every $h' \in H_j$ weakly following h that s_j leads to, is in $H_j^{bd}(s_j)$, it follows that s_j is optimal for type $\tau_j^{s_j}$ at every $h' \in H_j$ weakly following h that s_j leads to.

So, we have shown for every $s_j \in S_j^{bd}(h)$ that s_j is optimal for type $\tau_j^{s_j}$ at every $h' \in H_j$ weakly following h that s_j leads to. Since $b_i(t_i^{\sigma_i(s_i)}, h)$ only assigns positive probability to opponents' strategy-type pairs $(s_j, \tau_j^{s_j})$ where $s_j \in S_j^{bd}(h)$, we may conclude that type $t_i^{\sigma_i(s_i)}$ believes at h in the opponents' future rationality.

Hence, type $t_i^{\sigma_i(s_i)}$ believes in the opponents' future rationality.

Case 2. Consider now a type $\tau_i^{s_i}$ for some $s_i \in S_i$. We show that also type $\tau_i^{s_i}$ believes in the opponents' future rationality.

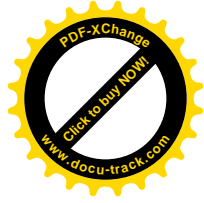
Take an arbitrary information set $h \in H_i$. Then, by construction, $b_i(\tau_i^{s_i}, h)$ only assigns positive probability to opponents' strategy-type pairs $(s_j, \tau_j^{s_j})$ with $s_j \in S_j^{bd}(h)$. From above, we know that every strategy $s_j \in S_j^{bd}(h)$ is optimal for type $\tau_j^{s_j}$ at every $h' \in H_j$ weakly following h that s_j leads to. This means, however, that type $\tau_i^{s_i}$ assigns at h only positive probability to opponents' strategy-type pairs $(s_j, \tau_j^{s_j})$ where s_j is optimal for type $\tau_j^{s_j}$ at every $h' \in H_j$ weakly following h that s_j leads to. That is, type $\tau_i^{s_i}$ believes at h in the opponents' future rationality.

Summarizing, we thus see that every type in the epistemic model believes in the opponents' future rationality. As a consequence, all types in the epistemic model express common belief in future rationality.

Step 5. $(\sigma_1(s_1), \dots, \sigma_n(s_n))$ leads to same outcome as (s_1, \dots, s_n) .

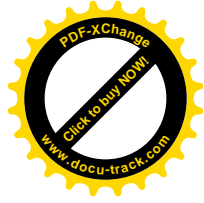
Finally, consider a strategy combination (s_1, \dots, s_n) in $\Gamma^{icd}(\emptyset)$ that leads to the outcome z . We show that the induced strategy combination $(\sigma_1(s_1), \dots, \sigma_n(s_n))$ leads to the same outcome z .

By construction, the strategy combination (s_1, \dots, s_n) only leads to information sets in H^{icd} . By Step 1, we know that at every information set $h \in H_i^{icd}$, the strategy $\sigma_i(s_i)$ prescribes the same choice as s_i . But



then, the strategy combination $(\sigma_1(s_1), \dots, \sigma_n(s_n))$ must lead to the same outcome z . This completes Step 5.

By using Steps 3, 4 and 5, it is now easy to prove Theorem 9.4.2. Consider, namely, an outcome z you initially deem possible under common strong belief in rationality. Then, there is a strategy combination (s_1, \dots, s_n) in $\Gamma^{icd}(\emptyset)$ that leads to the outcome z . By Step 5, we know that the induced strategy combination $(\sigma_1(s_1), \dots, \sigma_n(s_n))$ also leads to z . By Step 3 we know that every strategy $\sigma_i(s_i)$ is optimal for the type $t_i^{\sigma_i(s_i)}$ constructed in Step 2. Moreover, by Step 4, the type $t_i^{\sigma_i(s_i)}$ expresses common belief in future rationality. Hence, every strategy $\sigma_i(s_i)$ can rationally be chosen under common belief in future rationality. But then, the induced strategy combination $(\sigma_1(s_1), \dots, \sigma_n(s_n))$ is in $\Gamma^{bd}(\emptyset)$. As $(\sigma_1(s_1), \dots, \sigma_n(s_n))$ leads to z , it follows that also under common belief in future rationality, you initially deem possible the outcome z . This completes the proof of Theorem 9.4.2. ■



Practical Problems

9.1. Two parties in a row.

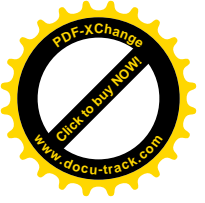
Recall the story from Problem 8.1 in Chapter 8.

- (a) What strategies can you rationally choose under common strong belief in rationality? Compare this to the strategies you can rationally choose under common belief in future rationality, found in Problem 8.1, part (c).
- (b) Under common strong belief in rationality, what colors can you rationally wear at the second party, provided you go to that party? Compare this to the colors you can rationally wear at the second party under common belief in future rationality, found in Problem 8.1, part (c). How do you explain this difference intuitively?
- (c) Find the unique self-confirming pair of rationality orderings.
- (d) Construct an epistemic model such that, for every strategy found in (a), there is a type that expresses common strong belief in rationality, and for which this strategy is optimal.

9.2. Selling ice cream.

Recall the story from Problem 8.2 in Chapter 8.

- (a) Which strategies can you and Barbara rationally choose under common strong belief in rationality? Compare this to the strategies you and Barbara can rationally choose under common belief in future rationality, found in Problem 8.2, part (b).
- (b) Suppose now that the weather forecast for tomorrow is more optimistic, such that you would expect the numbers in Figure 8.15 only to drop by 30% tomorrow. Which strategies can you and Barbara rationally choose under common strong belief in rationality? Again, compare this to the strategies you and Barbara can rationally choose under common belief in future rationality, found in Problem 8.2, part (d).
- (c) Consider the scenario in (b). Under common strong belief in rationality, what belief can Barbara hold *initially* about your strategy choice? And what beliefs can Barbara hold about your strategy choice if she observes that you will sell today?
- (d) For the scenario in (b), find the unique self-confirming pair of rationality orderings.
- (e) For the scenario in (b), construct an epistemic model such that, for every strategy found in (b), there is a type that expresses common strong



belief in rationality, and for which this strategy is optimal. Which types in your model express common strong belief in rationality? Which do not?

Hint for (e): Build an epistemic model with the following properties: For every $k \in \{0, 1, 2, \dots\}$ and every strategy s_i that can rationally be chosen under expressing up to k -fold strong belief in rationality, construct a type $t_i^{s_i}$ such that

- strategy s_i is optimal for type $t_i^{s_i}$, and
- type $t_i^{s_i}$ expresses up to k -fold strong belief in rationality.

For the construction of these types, use the self-confirming pair of rationality orderings found in (d).

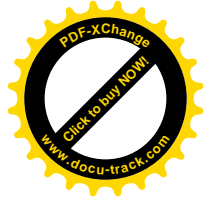
9.3. Watching TV with Barbara.

Recall the story from Example 9.2 in this chapter. So, before you both write down the program on a piece of paper, you have the option to start a fight with Barbara or not. Suppose now that, in case you decide *not* to start a fight with her, Barbara subsequently has the option to start a fight with *you* about the program to be watched. If Barbara indeed starts a fight with you then, similarly as before, this would reduce both your utility and Barbara's utility by 2.

- Model this situation as a dynamic game between you and Barbara.
- Find the strategies you can rationally choose under common strong belief in rationality. What about Barbara?
- Describe verbally the reasoning that leads to these strategy choices.
- What outcome do you initially expect under common strong belief in rationality? Compare this to the outcome you initially expected in the original situation of Example 9.2. Who has an advantage here under common strong belief in rationality, and why?
- Find the unique self-confirming pair of rationality orderings.
- Based on these rationality orderings, specify at each of your information sets what the possible beliefs are that you can hold about Barbara's strategy choice under common strong belief in rationality. Do the same for Barbara.

9.4. Never let a lady wait.

It is Saturday afternoon, and Barbara and you want to have dinner this evening at 8.00 pm. In the village where you live there are only two



restaurants – an Italian restaurant and a Chinese restaurant. The problem is that you prefer the Italian restaurant whereas Barbara prefers the Chinese restaurant. More precisely, having dinner in the Italian restaurant gives you a utility of 10, and eating in the Chinese restaurant yields you a utility of 7, but for Barbara it is the other way around. Even after a long discussion this morning you could not reach an agreement about the restaurant. For this reason you will both go to one of the restaurants this evening, without knowing to which restaurant the other person goes, and hope that you are lucky enough to find your friend there.

In the past you have built up a reputation of coming late, and Barbara knows that you have often used it as a strategic weapon. Also this evening you will strategically choose between arriving on time, or arriving one hour late. That is, at 7.45 pm you decide between walking to one of the restaurants and be there at 8.00 pm precisely, or wait until 8.45 pm to leave for one of the restaurants. In the latter case you would not know, however, in which restaurant Barbara is. Barbara, on the other hand, is always on time, and so it will be this evening. So, at 7.45 pm she decides to which restaurant she will go, and she will be there exactly at 8.00 pm. If you are both in the same restaurant at 8.00 pm then you will have dinner together, and the utilities will be as described above.

The other possibility is that Barbara, at 8.45 pm, is still waiting in front of an empty chair in her restaurant of choice. In that case, Barbara does not know whether you are momentarily at the other restaurant, or that you are waiting until 8.45 pm to leave your house. She then has two options – to stay in the same restaurant and hope that you will come at 9.00 pm, or to go to the other restaurant and hope that you will be there at 9.00 pm. On the other hand, you will never switch to the other restaurant if you left your house at 7.45 pm and are still alone in the restaurant at 8.45 pm – and Barbara knows this! If you are both in the same restaurant at 9.00 pm then you will have dinner together at 9.00 pm, but the utilities for both of you will be decreased by 2 units because of the one hour delay. Assume, moreover, that walking to the other restaurant would decrease Barbara's utility by 1 extra unit because it is raining outside. If you are both at different restaurants at 9.00 pm, then you will both be disappointed, go home and have a sandwich, yielding both of you a utility of 0 in total.

- (a) Model this situation as a dynamic game between you and Barbara. Be careful how to model the information sets!

- (b) Find the strategies that you and Barbara can rationally choose under common strong belief in rationality. Do you let Barbara wait?



	Italian	Chinese	Greek
You	3	2	1
Barbara	2	3	1
Chris	2	-1	3

Table 9.20: Utilities for you, Barbara and Chris in “Dinner for three”

- (c) Describe verbally the reasoning that leads to these strategy choices. Who has an advantage under common strong belief in rationality, and why?
- (d) Find the unique self-confirming pair of rationality orderings.
- (e) Based on these rationality orderings, specify at each of your information sets what the possible beliefs are that you can hold about Barbara’s strategy choice under common strong belief in rationality. Do the same for Barbara.

9.5. Dinner for three.

It is Saturday afternoon and you, Barbara and Chris would like to have dinner this evening. As in Problem 9.4, you have a strong preference for the Italian restaurant, whereas Barbara would rather go to the Chinese restaurant. Recently, Chris has discovered a new Greek restaurant in town which he likes very much, but you and Barbara are not very enthusiastic about it. Chris, on the other hand, is allergic to soy sauce, and he therefore would like to avoid Chinese food at all costs. Table 9.20 specifies the utilities that you, Barbara and Chris would derive from eating in each of the three restaurants. During the whole afternoon you have tried to reach an agreement about where to have dinner, but without success. Barbara therefore told you that she will be waiting in her favorite restaurant at 8.00 pm, and similarly Chris announced that he will be waiting in his favorite restaurant around that time. You told Barbara and Chris that you still have to make up your mind, and that at 8.00 pm you would decide to which of the three restaurants to go. At 8.15 pm Barbara will observe whether you have chosen her favorite restaurant or not. In both cases she may decide to stay, or to switch to one of the other two restaurants, which would take her around 15 minutes. Similarly for Chris. You, however, will stay where you are. If at 8.30 pm there are at least two friends in the same restaurant, then these friends will have dinner together. Anybody who is alone in the restaurant at 8.30 pm will go home, eat a sandwich, and go to bed early, yielding him or her a utility of 0.



(a) Model this situation as a dynamic game between you, Barbara and Chris. How many strategies does each of you have?

(b) Find the strategies that you, Barbara and Chris can rationally choose under common strong belief in rationality. Do you expect to have dinner with some of your friends? If so, with whom, and in which restaurant?

Hint for (b): The full decision problems in this game are very large, and therefore explicitly writing down these full decision problems is not a very good idea. Instead, use the graphical representation of the game you constructed in (a), and find a way to “eliminate strategies directly from the graphical representation”. This will save you a lot of writing.

(c) Describe verbally the reasoning that leads to these strategy choices. Who has an advantage under common strong belief in rationality, and why?

(d) Find the unique self-confirming combination of rationality orderings.

(e) Based on these rationality orderings, specify at each of Barbara's information sets what the possible beliefs are that she can hold about your strategy choice and Chris' strategy choice under common strong belief in rationality. Do the same for Chris.

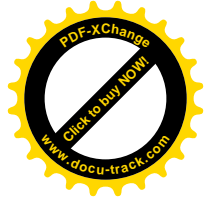
9.6. Read my mind.

You and Barbara participate in a TV show that is called “Read my mind”. The rules are very simple: On a table there are six different objects, which are numbered from 1 to 6. You must take one of the six objects, but Barbara cannot see this since she is blindfolded. However, the showmaster tells Barbara whether the number you chose is odd or even. Afterwards, Barbara must try to guess the number of the object you chose. If you choose object number k and Barbara guesses it correctly, then both you and Barbara get $1000k^2$ euros. If Barbara is wrong about the object you chose, you both get nothing.

(a) Model this situation as a game between you and Barbara.

(b) Find the strategies that you and Barbara can rationally choose under common strong belief in rationality. What outcomes do you deem possible under common strong belief in rationality?

Hint for (b): As in Problem 9.5, the full decision problems in this game are fairly large, and therefore explicitly writing down these full decision problems is not a very good idea. Instead, use the graphical representation of the game you constructed in (a), and find a way to



“eliminate strategies directly from the graphical representation”. This will save you a lot of writing.

- (c) Describe verbally the reasoning that leads to the strategy choices in (b).
- (d) Find the unique self-confirming pair of rationality orderings.
- (e) Construct an epistemic model such that, for every strategy found in (b), there is a type that expresses common strong belief in rationality, and for which this strategy is optimal.

Hint for (e): Build an epistemic model with the following properties: For every $k \in \{0, 1, 2, \dots\}$ and every strategy s_i that can rationally be chosen under expressing up to k -fold strong belief in rationality, construct a type $t_i^{s_i}$ such that

- strategy s_i is optimal for type $t_i^{s_i}$, and
- type $t_i^{s_i}$ expresses up to k -fold strong belief in rationality.

For the construction of these types, use the self-confirming pair of rationality orderings found in (d).

- (f) Find the strategies that you and Barbara can rationally choose under common belief in future rationality. What outcomes do you deem possible under common belief in future rationality? Compare this to your answers in (b).

9.7. Time to say goodbye.

After many years of friendship and numerous adventures, it is now time to say goodbye to your friends Barbara and Chris. They have decided to leave for another country, and this afternoon they both catch their flight at the local airport. A map of the airport can be found in Figure 9.4. As you can see, there are eight gates at the airport. Barbara’s flight is at 3.30 pm at gate 5, whereas Chris must catch his flight at 3.50 pm at gate 2, and everybody knows this. You have promised Barbara and Chris that you would be waiting at one of the eight gates at 3.00 pm to say goodbye, but you forgot to tell them at which gate. Barbara and Chris will be waiting at 3.00 pm at their respective gates of departure. It is now 2.00 pm, and you must decide at which gate you will be waiting.

Due to the shape of the airport, Barbara can only see you standing at your gate at 3.00 pm if you are waiting at gate 4, 5 or 6. Similarly, Chris can only see you standing if you have chosen gate 1, 2 or 3. In either case, Barbara and Chris have the option to leave for another gate at 3.00 pm in order to meet you, or to stay where they are. However,

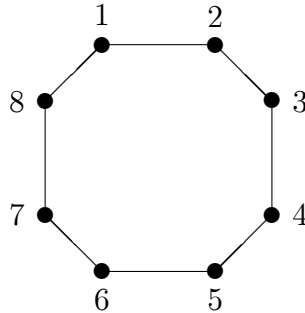
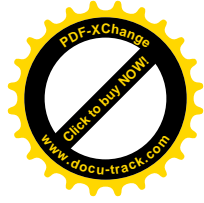


Figure 9.4: A map of the airport in “Time to say goodbye”

both must make sure that they will be back in time to catch their flight! Suppose that it takes 10 minutes to walk from one gate to the next gate.

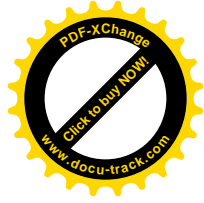
The utilities are as follows: If you meet only one of your friends at the gate, then your utility is the number of minutes you spend with that friend before he or she has to leave. If you meet both of your friends, then your utility is the sum of the numbers of minutes you spend with Barbara and Chris, plus a bonus of 40 for being all together. If you do not meet any of your friends, you will be very disappointed and your utility will be zero.

Barbara and Chris had a last fight yesterday evening, and are therefore only interested in seeing you. More precisely, the utility for Barbara and Chris is the number of minutes spent with you before leaving.

- (a) To which gates can Barbara and Chris walk at 3.00 pm?
- (b) Model this situation as a dynamic game between you, Barbara and Chris.
- (c) Find the strategies that you can rationally choose under common strong belief in rationality. How many friends do you expect to meet at the end?

Hint for (c): As in Problems 9.5 and 9.6, the full decision problems in this game are fairly large, and therefore explicitly writing down these full decision problems is not a very good idea. Instead, use the graphical representation of the game you constructed in (b), and find a way to “eliminate strategies directly from the graphical representation”. This will save you a lot of writing.

- (d) Describe verbally the reasoning that leads to the strategy choices in (c).



Theoretical Problems

9.8. Correct beliefs and common strong belief in rationality.

Consider a dynamic game with two players, say i and j . A type t_i for player i is said to *believe that j has correct beliefs* if at every information set $h \in H_i$, type t_i only assigns positive probability to types t_j for player j that, at every $h' \in H_j$, assign probability 1 to i 's type t_i . That is, t_i always believes that j is correct about i 's type, and hence about i 's beliefs.

(a) Show that, in general, it is impossible to construct a type t_i for player i that (1) expresses common strong belief in rationality, and (2) believes that j has correct beliefs.

Hint to (a): It is sufficient to take an example from Chapter 9, and show that in that particular example there is no type t_i for a certain player i that satisfies the two conditions above.

(b) For some of the concepts in this book, we were able to transform the concept into an associated “equilibrium concept” by additionally requiring that a player believes that his opponent is correct about his beliefs, and that a player believes that his opponent believes that he is correct about the opponent’s beliefs (in the case of two players). In this way, common belief in rationality could be transformed into Nash equilibrium (Chapter 4), common full belief in “caution and primary belief in rationality” could be transformed into perfect equilibrium (Problem 5.9), and common full belief in “caution and respect of preferences” could be transformed into proper equilibrium (Problem 6.9). Can we transform the concept of common strong belief in rationality into an associated equilibrium concept? Explain your answer.

9.9. Initial belief in the opponents’ rationality.

Suppose that we extend the epistemic model for a dynamic game in the following way: Not only does a type t_i hold a conditional belief about the opponents’ strategy-type combinations at each of his information sets $h \in H_i$, but *also* at the beginning of the game \emptyset . That is, before the game starts every type t_i holds an *initial* belief $b_i(t_i, \emptyset) \in \Delta(S_{-i} \times T_{-i})$ about the opponents’ strategy choices and beliefs. Here, S_{-i} denotes the set of opponents’ strategy combinations, and T_{-i} the set of opponents’ type combinations.

In this extended epistemic model, say that a strategy s_i is *optimal* for type t_i if (1) s_i is optimal for t_i 's *initial* belief about the opponents’ strategy choices, and (2) at every information set $h \in H_i$ that s_i leads



to, strategy s_i is optimal for t_i 's conditional belief at h about the opponents' strategy choices. Moreover, type t_i is said to *initially believe in his opponents' rationality* if t_i 's initial belief $b_i(t_i, \emptyset)$ only assigns positive probability to opponents' strategy-type pairs (s_j, t_j) where the strategy s_j is optimal for the type t_j . We say that type t_i expresses *common belief in "initial belief in rationality"* if

- t_i initially believes in his opponents' rationality,
- t_i assigns, at \emptyset and at every $h \in H_i$, only positive probability to opponents' types that initially believe in their opponents' rationality,
- t_i assigns, at \emptyset and at every $h \in H_i$, only positive probability to opponents' types t_j that assign, at \emptyset and at every $h' \in H_j$, only positive probability to opponents' types that initially believe in their opponents' rationality,

and so on.

Finally, a strategy s_i can rationally be chosen under common belief in "initial belief in rationality" if there is an extended epistemic model, and a type t_i within it, such that strategy s_i is optimal for t_i , and type t_i expresses common belief in "initial belief in rationality".

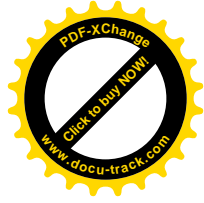
(a) Construct an algorithm, similar to those in Chapters 8 and 9, that selects for every player precisely those strategies he can rationally choose under common belief in "initial belief in rationality".

(b) Show that every strategy that can rationally be chosen under common belief in future rationality, can also rationally be chosen under common belief in "initial belief in rationality".

(c) Give an example of a dynamic game in which there is an outcome z such that (1) you can initially deem outcome z possible under common belief in "initial belief in rationality", but (2) you never initially deem outcome z possible under common belief in future rationality.

(d) Show that every strategy that can rationally be chosen under common strong belief in rationality, can also rationally be chosen under common belief in "initial belief in rationality".

(e) Give an example of a dynamic game in which there is an outcome z such that (1) you can initially deem outcome z possible under common belief in "initial belief in rationality", but (2) you never initially deem outcome z possible under common strong belief in rationality.

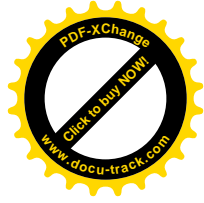
**9.10. Property of the iterated conditional dominance procedure.**

Say that an information set h is initially deemed possible under common strong belief in rationality, if for every player i there is a strategy s_i that can rationally be chosen under common strong belief in rationality, such that this combination of strategies leads to h .

For every player i and every information set $h \in H_i$, let $\Gamma^{icd}(h) = (S_i^{icd}(h), S_{-i}^{icd}(h))$ be the reduced decision problem at h that remains after applying the iterated conditional dominance procedure. Let $S_i^{icd}(\emptyset)$ be the set of strategies for player i that survive the iterated conditional dominance procedure at \emptyset . Finally, let $S_i(h)$ be the set of strategies for player i that lead to h , and let $S_{-i}(h)$ be the set of opponents' strategy combinations that lead to h .

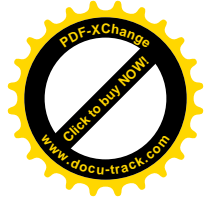
(a) Show that an information set $h \in H_i$ is initially deemed possible under common strong belief in rationality, if and only if, $S_i^{icd}(h) = S_i^{icd}(\emptyset) \cap S_i(h)$ and $S_{-i}^{icd}(h) = S_{-i}^{icd}(\emptyset) \cap S_{-i}(h)$.

(b) Does the property in (a) also hold for the concept of common belief in future rationality, and the associated backward dominance procedure? If so, then provide a proof. If not, then provide a counterexample.



Literature

Common strong belief in rationality. The main idea we explored in this chapter is that of *strong belief in the opponents' rationality*, which states that, whenever it is possible for you to believe at an information set that your opponents' have chosen rationally, then you *must* believe at that information set that your opponents have chosen rationally. This idea has first been formalized by Battigalli and Siniscalchi (2002). By iterating this condition, Battigalli and Siniscalchi developed a concept which we have called *common strong belief in rationality*. Compared to our approach in this chapter, there are two major differences with the construction by Battigalli and Siniscalchi. First, Battigalli and Siniscalchi use an epistemic model for dynamic games in which it is assumed that the players' conditional beliefs satisfy *Bayesian updating*, whereas we do not. More precisely, Battigalli and Siniscalchi use *conditional probability systems* (see also the literature section of Chapter 5) as the basic ingredient of their epistemic model, which assumes that the conditional beliefs satisfy Bayesian updating whenever possible. More than this, the epistemic model they use actually assumes that the players express *common belief in Bayesian updating*. Secondly, Battigalli and Siniscalchi use a *complete* and *terminal* epistemic model (see the literature section of Chapter 8) for dynamic games, which contains all possible belief hierarchies one can think of – provided they express common belief in Bayesian updating – whereas we do not require the epistemic model to be complete. In fact, for our purposes in this book it is sufficient to work with *finite* epistemic models, containing only finitely many belief hierarchies – or types – for each player. What we do need, however, is that the epistemic model contains *sufficiently many* types. Remember that, in order to properly define *strong belief in the opponents' rationality*, we need the following “richness” condition on the epistemic model M : If at a given information set h for player i there is an opponents' combination of types – possibly outside M – for which there is a combination of optimal strategies leading to h , then the epistemic model M must contain at least one such opponents' combination of types for which there is a combination of optimal strategies leading to h . Battigalli and Siniscalchi guarantee this “richness” condition by insisting on an epistemic model M that contains *all* possible belief hierarchies – or types – which satisfy common belief in Bayesian updating, that is, by using epistemic models M which are terminal. Similarly, the “richness” conditions which we need to define k -fold strong belief in rationality, for $k \geq 2$, are automatically satisfied by Battigalli and Siniscalchi's model, as they assume the epistemic model to be complete and terminal.

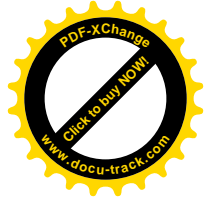


Extensive form rationalizability. The idea of common strong belief in rationality is already implicitly present in the concept of *extensive form rationalizability*, as proposed by Pearce (1984). In fact, Pearce (1984) presented the concept of extensive form rationalizability by means of an iterated elimination procedure, which recursively eliminates strategies and conditional belief vectors from the game. Later, Battigalli (1997) presented a simpler elimination procedure, which he proved to be equivalent to Pearce's original procedure.

The procedures by Pearce and Battigalli can be summarized as follows: For every player i , let S_i^0 be the set of all strategies, and let B_i^0 be the set of all conditional belief vectors b_i about the opponents' strategies, satisfying Bayesian updating. At every step $k \geq 1$, let B_i^k be the set of conditional belief vectors $b_i = (b_i(h))_{h \in H_i}$ from B_i^{k-1} satisfying the following condition: If at information set $h \in H_i$ there is some opponents' strategy combination in S_{-i}^{k-1} leading to h , then $b_i(h)$ must assign positive probability only to opponents' strategy combinations in S_{-i}^{k-1} leading to h . Let S_i^k be the set of strategies s_i for player i that are optimal, at every information set $h \in H_i$ that s_i leads to, for some conditional belief vector b_i in B_i^k . A strategy s_i is then called *extensive form rationalizable* if s_i is in S_i^k for every k .

Note that the inductive step, which selects from B_i^{k-1} those conditional belief vectors which belong to B_i^k , is very similar in spirit to the condition of k -fold strong belief in rationality. Indeed, Battigalli and Siniscalchi (2002) show in their Proposition 6 that the strategies that can rationally be chosen under common strong belief in rationality, are exactly the extensive form rationalizable strategies in the sense of Pearce (1984) and Battigalli (1997). That is, common strong belief in rationality can be viewed as an epistemic foundation for the procedure of extensive form rationalizability.

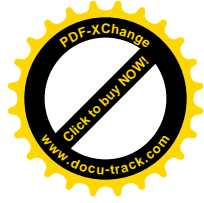
Algorithm. In this chapter we have presented an algorithm – the *iterated conditional dominance procedure* – which yields precisely those strategies that can rationally be chosen under common strong belief in rationality. The algorithm is due to Shimoji and Watson (1998), who show that their algorithm yields exactly the extensive form rationalizable strategies in a game. If we combine this result with Battigalli and Siniscalchi's theorem, showing that the extensive form rationalizable strategies are exactly the strategies that can rationally be chosen under common strong belief in rationality, we obtain Theorem 9.3.3 in this chapter, which shows that the iterated conditional dominance procedure generates precisely those strategies that can rationally be chosen under common strong belief in rationality.



Order (in)dependence. We have seen that for the iterated conditional dominance procedure, the order of elimination crucially matters for the strategies that survive the algorithm. That is, if at some step of the algorithm we either do not eliminate all strategies we can at a given information set, or do not scan through all information sets, then we may end up with different sets of strategies for the players at the end. This is in contrast with the backward dominance procedure, where the order of elimination is not relevant for the strategies that survive the algorithm. Chen and Micali (2011) and Robles (2006) show, however, that the order of elimination in the iterated conditional dominance procedure does not affect the *outcomes* that can be reached. More precisely, if we change the order of elimination in the iterated conditional dominance procedure, then we may change the sets of strategies that survive for the various players in the game, but the *outcomes* that can be reached by the surviving strategy combinations will remain the same!

Comparison with common belief in future rationality. In Section 9.4 of this chapter we have compared the concept of common strong belief in rationality with the concept of common belief in future rationality. And in particular we have compared the two associated algorithms with each other – the iterated conditional dominance procedure and the backward dominance procedure. These comparisons are largely based on Perea (2010). Among other things, we have shown that in terms of *outcomes* the concept of common strong belief in rationality is more restrictive than the concept of common belief in future rationality. More precisely, we prove in Theorem 9.4.2 that every outcome which is initially deemed possible under common strong belief in rationality, is also initially deemed possible under common belief in future rationality.

This theorem can also be proven by making use of the result by Chen and Micali (2011) and Robles (2006) above, stating that the *outcomes* that can be reached under the strategies that survive the iterated conditional dominance procedure, are independent of the order of elimination. Namely, it can be shown that the backward dominance procedure corresponds to the first few steps of the iterated conditional dominance procedure – by eliminating strategies in a very particular, different order – but *without completing* the elimination procedure. Since we know, by the result of Chen and Micali (2011) and Robles (2006), that the order of elimination in the iterated conditional dominance procedure does not change the outcomes that can be reached, it follows that the outcomes that can be reached under the iterated conditional dominance procedure will be a subset of the outcomes that can be reached under the backward dominance procedure. But then, the outcomes that are initially deemed



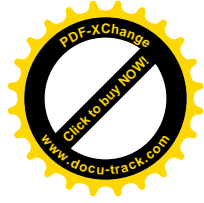
possible under common strong belief in rationality will be a subset of the outcomes that are initially deemed possible under common belief in future rationality. Hence, the result in Theorem 9.4.2 follows from this argument.

Common strong belief in rationality and backward induction.

In Corollary 9.4.3 we have shown that in every game with perfect information, every outcome that is initially deemed possible under common strong belief in rationality must be a backward induction outcome in that game. We have shown this result by the following steps: First, we use the result mentioned above, stating that every outcome which is initially deemed possible under common strong belief in rationality, is also initially deemed possible under common belief in future rationality. Moreover, we use the insight that in games with perfect information, the backward dominance procedure – and hence the concept of common belief in future rationality – is equivalent to the backward induction procedure. By combining these two pieces we conclude that in games with perfect information, every outcome which is initially deemed possible under common strong belief in rationality, must be an outcome that is reachable under the backward induction procedure – that is, must be a backward induction outcome.

This result has first been proved by Battigalli (1997) – for the case of *generic* games with perfect information – in his Theorem 4. Here, a game with perfect information is called *generic* if for a given player, all terminal histories yield different utilities. In fact, Battigalli (1997) shows that in every generic game with perfect information, there is only one outcome that can be reached under extensive form rationalizability, namely the unique backward induction outcome in that game. The proof that Battigalli delivers is very different from ours, however. He uses properties of fully stable sets by Kohlberg and Mertens (1986) to show the result, whereas we use a more direct and constructive way to prove the statement.

Gradwohl and Heifetz (2011) investigate *non-generic* games with perfect information – where different terminal histories may induce the same utility for a player – and compare the concepts of common strong belief in rationality (both in the usual sense and for the agent form), common belief in future rationality and subgame perfect equilibrium for such games. Here, the *agent form* is the game that is obtained by assigning a different player to every information set. They find that for these games, in terms of outcomes reached, subgame perfect equilibrium is more restrictive than common strong belief in rationality, which is more restrictive



than common belief in future rationality, which in turn is more restrictive than common strong belief in rationality for the agent form.

Rationality orderings. The concept of *rationality orderings*, as discussed in Section 9.6, has been introduced and analyzed by Battigalli (1996). The term *self-confirming combination of rationality orderings* we use, corresponds to what Battigalli calls *correlated sequential rationality orderings*. The main idea behind rationality orderings is that a player, at each of his information sets, always looks for the “most rational” opponents’ strategies that could have led to this information set. Battigalli (1996) calls this the *best rationalization principle*. The theorem in Section 9.6, which states that there is unique self-confirming combination of rationality orderings – namely the one induced by common strong belief in rationality – is based on Battigalli’s Theorem 2.2.

Bayesian updating. In Theorem 9.7.1 we have stated that for the concept of common strong belief in rationality, it is not relevant whether we additionally impose (common belief in) Bayesian updating or not. For the eventual strategy choices selected by the concept it does not make a difference. This property has first been shown by Shimoji and Watson (1998) in their Theorem 2. More precisely, they show that for the strategy choices selected by the iterated conditional dominance procedure, it is inessential whether we impose (common belief in) Bayesian updating or not. As the iterated conditional dominance procedure delivers exactly the strategies that can rationally be chosen under common strong belief in rationality, the result in Theorem 9.7.1 follows.

Forward induction. The concept of common strong belief in rationality – or equivalently, the concept of extensive form rationalizability – is considered to be a *forward induction* concept. In the literature there is no unique definition of *forward induction*, but intuitively it describes a way of reasoning in which a player, when confronted with an unexpected opponent’s choice, tries to find a “most plausible” explanation for this surprising move by his opponent.

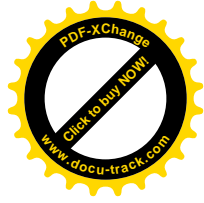
Indeed, common strong belief in rationality is perfectly in line with such a way of reasoning. According to this concept, a player asks at each of his information sets whether there are optimal opponents’ strategies that could have led to this information set. If so, he will ascribe positive probability only to optimal opponents’ strategies at that information set. He then asks whether there are optimal opponents’ strategies, leading to his information set, which can rationally be chosen by opponents that strongly believe in their opponents’ rationality. If so, then he will ascribe positive probability only to such opponents’ strategies that are optimal



for opponents that strongly believe in their opponents' rationality. And so on. So, the player looks for the "most plausible" opponents' strategies that could have led to his information set, and assigns positive probability only to such "most plausible" opponents' strategies. It therefore matches precisely our intuition of forward induction reasoning.

In the literature, most forward induction concepts presented so far are *equilibrium concepts*. More precisely, most of these concepts take the notion of *sequential equilibrium* as defined by Kreps and Wilson (1982) – see the literature section of Chapter 8 for a discussion – and impose additional restrictions on the beliefs about opponents' *past* choices within this concept. Examples of this kind are *forward induction equilibrium* (Cho (1987)), *justifiable sequential equilibrium* (McLennan (1985)) and *stable sets of beliefs* (Hillas (1994)) for general dynamic games, and the *intuitive criterion* (Cho and Kreps (1987)) and its various refinements for the special class of signaling games. See Chapter 5 in Perea (2001) for an overview of most of these forward induction refinements of sequential equilibrium.

What is a bit problematic about these forward induction refinements of sequential equilibrium is that they are incapsulated within a *backward induction* concept – namely *sequential equilibrium*. In the literature section of Chapter 8, namely, we have seen that the concept of sequential equilibrium implicitly assumes *common belief in future rationality*, which is a typical *backward induction* type of reasoning, as it requires players to only critically reason about the choices that opponents will make now and in the future. So, in a sense, the forward induction refinements of sequential equilibrium are a mix of backward induction and forward induction arguments, which makes the forward induction reasoning less transparent as it could be. Moreover, in some games such concepts will be unable to filter out the "natural" forward induction strategy for a player. Consider, for instance, the example "Painting Chris' house" which we discussed at the beginning of this chapter. In that game, the "natural" forward induction strategy for you is to choose a price of 300. Namely, if you observe that Barbara has rejected the colleague's offer, this can only be part of a rational strategy for Barbara if she subsequently will choose a price of 400. Hence, in the spirit of forward induction, it makes sense for you to believe that Barbara will indeed choose a price of 400 if you observe that Barbara has rejected the colleague's offer. But then, your only optimal choice is to choose a price of 300 yourself. However, a forward induction refinement of sequential equilibrium will never be able to select this price of 300 for you! The reason is that there is only one sequential equilibrium in this game, in which you believe – after observing that Barbara has rejected the colleague's offer – that Barbara will



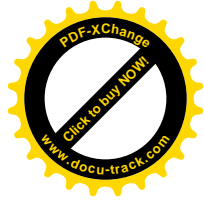
choose a price of 200. Hence, your only optimal choice under a sequential equilibrium is to choose a price of 200 as well. That is, every forward induction refinement of sequential equilibrium will uniquely select the choice 200 for you, which is not the “natural” forward induction choice for you in this game.

In contrast, the concept of *common strong belief in rationality* is a “pure” forward induction concept, which is not encapsulated within any backward induction concept like common belief in future rationality. If we were to build a forward induction concept which takes the backward induction concept of common belief in future rationality as a starting point, then within the example “Painting Chris’ house” we would never be able to select your “natural” forward induction choice 300, as under common belief in future rationality there is only one strategy you can rationally choose, namely 200.

Explicable equilibrium. The forward induction concept of *explicable equilibrium*, as proposed by Reny (1992b), is different from the concepts discussed above, as it does not take sequential equilibrium as a starting point, but rather the weaker concept of *weak sequential rationality*. Remember from the literature section in Chapter 8 that *weak sequential rationality* intuitively means that players *initially* believe in the opponents’ future rationality, but do not necessarily believe so at later stages of the game. This is weaker than the *sequential rationality* condition imposed in sequential equilibrium, which states that players *always* believe in their opponents’ future rationality.

The concept of *explicable equilibrium* is very similar – in spirit – to the concept of *extensive form rationalizability*, and equivalently the concept of *common strong belief in rationality*. It also defines a rationality ordering on the players’ sets of strategies, and imposes that a player, at each of his information sets, looks for the “most rational” opponents’ strategies that could have led to this information set, and assigns positive probability only to such “most rational” opponents’ strategies. However, the rationality orderings are defined differently than under common strong belief in rationality, as the concept of explicable equilibrium takes weakly sequentially rational assessments as a starting point for the analysis.

Reny (1992b) shows, in his Proposition 3, that for *generic* games with perfect information, the concept of explicable equilibrium always leads to the unique *backward induction outcome*, but not necessarily to the unique backward induction strategies for the players. This result is similar to Theorem 4 in Battigalli (1997), which shows that the same holds for the concept of extensive form rationalizability – and hence also for common strong belief in rationality. Like Battigalli, also Reny proves



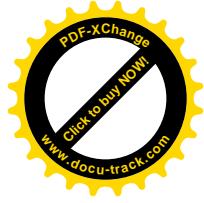
this result by making use of properties of fully stable sets as defined by Kohlberg and Mertens (1986).

Burning money games. The example “Watching TV with Barbara” which we discussed in this chapter, can be viewed as an instance of a *burning money game* – a type of game first studied by van Damme (1989) and Ben-Porath and Dekel (1992). A *burning money game* is a game with two players in which player 1, at the beginning of the game, can publicly and voluntarily burn a certain amount of money – and thereby reduce his utility – before facing a simultaneous move game with player 2. In the example “Watching TV with Barbara” you have the option to start a fight with Barbara at the beginning of the game, which would reduce your utility by 2. This can thus be seen as “publicly burning 2 utility units” before facing the game with Barbara.

In his paper, van Damme (1989) studies a burning money game in which player 1 can only burn one given amount of money, whereas Ben-Porath and Dekel (1992) analyze games in which player 1 can freely choose the amount of money he would like to burn. Both papers use the algorithm of *iterated elimination of weakly dominated strategies* – as studied in Chapter 7 of this book – to analyze the burning money game. More precisely, they take the full decision problem at the beginning of the game, and iteratedly remove all weakly dominated strategies from that decision problem.

Ben-Porath and Dekel (1992) show the following striking result: Suppose that in the simultaneous move game between players 1 and 2, there is a combination of choices (c_1^*, c_2^*) which (1) is strictly better for player 1 than any other choice-combination in that game, and (2) is strictly better for player 2 than any other choice-combination (c_1^*, c_2) that involves c_1^* for player 1. Then, iterated elimination of weakly dominated strategies leads to a unique outcome, where player 1 does not burn any money, and player 1’s most preferred choice-combination (c_1^*, c_2^*) is obtained in the game that follows. This is similar to our finding in the example “Watching TV with Barbara”, where under common strong belief in rationality you will not start a fight with Barbara, and expect to watch your favorite program together with Barbara.

Later, Shimoji (2002) has shown that the result above by Ben-Porath and Dekel (1992) also holds if we use *extensive form rationalizability* – or equivalently *common strong belief in rationality* – instead of iterated elimination of weakly dominated strategies. That is, in the burning money games studied in Ben-Porath and Dekel (1992), also common strong belief in rationality always leads to a unique outcome, in which you do not



burn any money, yet receive your most preferred outcome in the game that follows.

In particular, in the burning money games studied by Ben-Porath and Dekel (1992), the algorithm of iterated elimination of weakly dominated strategies always yields the same result as the concept of common strong belief in rationality. This is actually true for other dynamic games of interest as well, which is the reason that people have often used the algorithm of iterated elimination of weakly dominated strategies as a *forward induction concept* for such dynamic games.