Games and Economic Behavior 83 (2014) 231–254

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Belief in the opponents' future rationality ☆

Andrés Perea

Maastricht University, EpiCenter & Department of Quantitative Economics, P.O. Box 616, 6200 MD Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 5 August 2011
Available online 27 November 2013

JEL classification:
C72

Keywords:
Epistemic game theory
Dynamic games
Belief in future rationality
Backwards rationalizability procedure
Backward dominance procedure
Backward induction

ABSTRACT

For dynamic games we consider the idea that a player, at every stage of the game, will always believe that his opponents will choose rationally in the future. This is the basis for the concept of *common belief in future rationality*, which we formalize within an epistemic model. We present an iterative procedure, *backward dominance*, that proceeds by eliminating strategies from the game, based on strict dominance arguments. We show that the backward dominance procedure selects precisely those strategies that can rationally be chosen under common belief in future rationality if we would not impose (common belief in) Bayesian updating.

1. Introduction

The goal of *epistemic game theory* is to describe plausible ways in which a player may reason about his opponents before he makes a decision in a game. In static games, the epistemic program is largely based upon the idea of *common belief in rationality* (Tan and Werlang, 1988), which states that a player believes that his opponents choose rationally, believes that every opponent believes that each of his opponents chooses rationally, and so on.

Extending this idea to dynamic games, however, does not come without problems. One obstacle is that in dynamic games it may be impossible to require that a player always believes that his opponents have chosen rationally *in the past*. Consider, for instance, a two-player game where player 1, at the beginning of the game, can choose between stopping the game and entering a subgame with player 2. If player 1 stops the game he would receive a utility of 10, whereas entering the subgame would always give him a lower utility. If player 2 observes that player 1 has entered the subgame, he cannot believe that player 1 has chosen rationally in the past.[1] In particular, it will not be possible in this game to require that player 2 always believes that player 1 chooses rationally *at all points in time*. In dynamic games, we are therefore forced to weaken the notion of *belief in the opponent's rationality*. But how?

In this paper we present one such way. We require that a player, under all circumstances, believes that his opponents will *choose rationally now and in the future*. So, even if a player observes that an opponent has chosen irrationally in the past, this should not be a reason to drop his belief in this opponent's present and future rationality. In order to keep our terminology short, we refer to this condition as *belief in the opponent's future rationality*, so we omit belief in *present* rationality in this

☆ I would like to thank Christian Bach, Pierpaolo Battigalli, Aviad Heifetz, an associate editor, and some anonymous referees for useful comments and suggestions.

E-mail address: a.perea@maastrichtuniversity.nl.
URL: http://www.personeel.unimaas.nl/a.perea/.

[1] At least, if we stick to a framework with complete information in which the players' utility functions are transparent to everyone, as we assume in this paper.

0899-8256/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.geb.2013.11.008

phrase. The reader should bear in mind, however, that we always assume belief in the opponent's *present* rationality as well. A first observation is that belief in the opponents' future rationality is always possible: Even if an opponent has behaved irrationally in the past, it is always possible to believe that he will choose optimally now and at all future instances.

Belief in the opponents' future rationality is certainly not the only reasonable condition one can impose on a player's beliefs in a dynamic game, but we think it provides a natural and plausible way of reasoning about the opponents. In a sense, it assumes that the player is *completely forward looking* – he only reasons about the opponents' behavior in the future of the game, and takes the opponents' past choices for granted without drawing any conclusions from these. A possible explanation the player could give for unexpected past choices is that his opponents were making mistakes, or misjudging the situation at hand, but this should, according to the concept of belief in the opponents' future rationality, not be a reason to give up the belief that these opponents will choose rationally in the future. This condition can thus be viewed as a typical *backward induction* condition, as opposed to *forward induction* reasoning which assumes that the player, at every stage of the game, tries to interpret the opponents' past choices as being part of some rational plan. There is something to say for both lines of reasoning, but in this paper we focus on the first one.

In this paper, we do not only impose that a player always believes in his opponents' future rationality, we also require that a player always believes that every opponent always believes in his opponents' future rationality, and that he always believes that every opponent always believes that every other player always believes in his opponents' future rationality, and so on. This leads to the concept of *common belief in future rationality*, which is the central idea in this paper.

As a first step, we lay out a formal epistemic model for finite dynamic games with complete information, and formalize the notion of common belief in future rationality within this model. This enables us to define precisely which strategies can be chosen by every player under common belief in future rationality. It turns out that the strategies that can rationally be chosen under common belief in future rationality are precisely the strategies that survive Penta's (2009) *backwards rationalizability procedure* – a procedure that iteratively eliminates strategies and conditional belief vectors from the game.

We go on by presenting a *new* iterative procedure – which we call the *backward dominance procedure* – that solely eliminates strategies (not conditional belief vectors) from the game, based on strict dominance arguments. The procedure works as follows: In the first round we eliminate, at every information set, those strategies for player i that are strictly dominated at a present or future information set for player i. In every further round k we eliminate, at every information set, those strategies for player i that are strictly dominated at a present or future information set h_i for player i, given the opponents' strategies that have survived until round k at that information set h_i. We continue until we cannot eliminate anything more. The strategies that eventually survive at the beginning of the game are those that survive the backward dominance procedure.

We show that the strategies that survive the backward dominance procedure are exactly the strategies that can rationally be chosen under common belief in future rationality if we would *not impose* (*common belief in*) *Bayesian updating*. In fact, imposing (common belief in) Bayesian updating can matter for the strategies that can eventually be chosen under common belief in future rationality. Consequently, the backward dominance procedure will in general select a set of strategies that *contains* the strategies that can rationally be chosen under common belief in future rationality, but it can actually contain more. In that sense, the backward dominance procedure can be used as a first selection towards the precise set of strategies that corresponds to common belief in future rationality. This is an important result, since the backward dominance procedure is particularly easy to use.

If we apply the backwards rationalizability procedure – or even the backward dominance procedure – to games with perfect information, then we obtain precisely the well-known backward induction procedure. As a consequence, applying common belief in future rationality to games with perfect information leads to backward induction.

The idea of (common) belief in the opponents' future rationality is not entirely new. For games with perfect information, some variants of it have served as an epistemic foundation for backward induction. See, for instance, Asheim (2002), Baltag et al. (2009), Feinberg (2005) and Samet (1996). In fact, the condition of "stable belief in dynamic rationality" in Baltag et al. (2009) matches exactly our definition of belief in the opponents' future rationality, although they restrict to a non-probabilistic framework. The reader may consult Perea (2007) for a detailed overview of the various epistemic foundations that have been offered for backward induction in the literature.

For general dynamic games, belief in the opponents' future rationality is *implicitly* present in "backward induction concepts" such as sequential equilibrium (Kreps and Wilson, 1982) and sequential rationalizability (Dekel et al., 1999, 2002 and Asheim and Perea, 2005). In fact, we show in Section 7 that sequential equilibrium and sequential rationalizability are both more restrictive than common belief in future rationality.

Now, why should we be interested in common belief in future rationality as a concept if it is already implied by sequential equilibrium and sequential rationalizability? We believe there are several important reasons.

First, the concept of common belief in future rationality is based upon very elementary decision theoretic and epistemic conditions, namely that players should always believe that their opponents will choose rationally in the remainder of the game, and that there is common belief throughout the game in this event. No other conditions, besides (common belief in) Bayesian updating, are imposed. In particular, we impose no equilibrium conditions as in sequential equilibrium. So, in this sense, common belief in future rationality constitutes a very basic concept. Compared to sequential rationalizability, the concept of common belief in future rationality is very explicit about the epistemic assumptions being made. In the formulation of sequential rationalizability, the epistemic conditions imposed are somewhat more hidden in the various ingredients of its definition.

Second, the concept of sequential equilibrium may rule out reasonable choices in some games, precisely because it imposes equilibrium conditions which are hard to justify if the game is played only once, and the players cannot communicate before the game. See Bernheim (1984) for an early and similar critique to Nash equilibrium.

Finally, we provide some iterative procedures that support the concept of common belief in future rationality, making the concept attractive also from a practical point of view. In general, sequential equilibrium strategies are much harder to compute.

In Section 7 we also compare our notion with the concept of *extensive form rationalizability* (Pearce, 1984; Battigalli, 1997; Battigalli and Siniscalchi, 2002) and find that, in terms of strategy choices, there is no general logical relationship between the two. In fact, there are games where both notions provide a unique, but different, strategy choice for a player. However, in terms of *outcomes* that can be reached, extensive form rationalizability is more restrictive than common belief in future rationality. Namely, every outcome that can be reached under extensive form rationalizability can also be reached under common belief in future rationality, but not vice versa. The reader is referred to Chapter 9 in Perea (2012) for a formal statement and proof of this result. Moreover, in Section 7 we compare our backward dominance procedure with the *iterated conditional dominance procedure* (Shimoji and Watson, 1998), which characterizes extensive form rationalizability. Both algorithms are similar in spirit, as they proceed by successively eliminating strategies at every information set in the game. However, the criteria that are used to eliminate a strategy at a given information set are different in both procedures. In Section 7 we precisely describe the differences and similarities between the two procedures.

The outline of this paper is as follows. In Section 2 we give some basic definitions and introduce an epistemic model for dynamic games. In Section 3 we formalize the idea of common belief in future rationality within this epistemic model. In Section 4 we present Penta's (2009) backwards rationalizability procedure, and show that it yields precisely those strategies that can rationally be chosen under common belief in future rationality. In Section 5 we introduce the backward dominance procedure, and show that the procedure selects exactly those strategies that can rationally be chosen under common belief in future rationality if we would not impose (common belief in) Bayesian updating. In Section 6 we discuss some important properties of the concept of common belief in future rationality and the associated procedures. In Section 7 we explore the relation between common belief in future rationality and other concepts for dynamic games such as sequential rationalizability and extensive form rationalizability. In Section 8 we discuss possible lines for future research. Section 9 contains all the proofs.

2. Model

In this section we formally present the class of dynamic games we consider, and explain how to build an epistemic model for such dynamic games.

2.1. Dynamic games

In this paper we restrict attention to dynamic games with *complete information*, in which the players' utility functions are transparent to everyone. By I we denote the set of players, by X the set of non-terminal histories (or nodes) and by Z the set of terminal histories. By $\emptyset$ we denote the beginning (or root) of the game. For every player i, we denote by H_i the collection of information sets for that player. Every information set $h_i \in H_i$ consists of a set of non-terminal histories. At every information set $h_i \in H_i$, we denote by $C_i(h_i)$ the set of choices (or actions) for player i at h_i. We assume that all sets mentioned above are *finite*, and hence we restrict to *finite* dynamic games in this paper. Finally, for every terminal history z and player i, we denote by $u_i(z)$ the utility for player i at z. As usual, we assume that there is *perfect recall*, meaning that a player never forgets what he previously did, and what he previously knew about the opponents' past choices.

We explicitly allow for *simultaneous moves* in the dynamic game. That is, we allow for non-terminal histories at which several players make a choice. Formally, this means that for some non-terminal histories x there may be different players i and j, and information sets $h_i \in H_i$ and $h_j \in H_j$, such that $x \in h_i$ and $x \in h_j$. In this case, we say that the information sets h_i and h_j are *simultaneous*. Explicitly allowing for simultaneous moves is important in this paper, especially for describing the concept of *common belief in future rationality*.

Say that an information set h *follows* some other information set h' if there are histories $x \in h$ and $y \in h'$ such that y is on the unique path from the root to x. Finally, we say that information set h *weakly follows* h' if either h follows h', or h and h' are simultaneous. We write $h \succcurlyeq h'$ to denote that h weakly follows h'.

We assume, throughout this paper, that there is an *unambiguous ordering of the information sets* in the game. That is, if information set h follows information set h', then h' does not follow h. Or, equivalently, there cannot be histories $x, y \in h$, and histories $x', y' \in h'$ such that x is on the path from the root to x', and y' is on the path from the root to y. This will be important for the concept of common belief in future rationality that we will develop.

To illustrate the concepts defined above, let us have a look at the example in Fig. 1. At the beginning of the game, $\emptyset$, player 1 chooses between a and b, and player 2 simultaneously chooses between c and d. So, $\emptyset$ is an information set that belongs to both players 1 and 2. If player 1 chooses b, the game ends, and the utilities are as depicted. If he chooses a, then the game moves to information set $h_{2.1}$ or information set $h_{2.2}$, depending on whether player 2 has chosen c or d. Player 1, however, does not know whether player 2 has chosen c or d, so player 1 faces information set h_1 after choosing a. Hence, $h_{2.1}$ and $h_{2.2}$ are information sets that belong only to player 2, whereas h_1 is an information set that belongs only

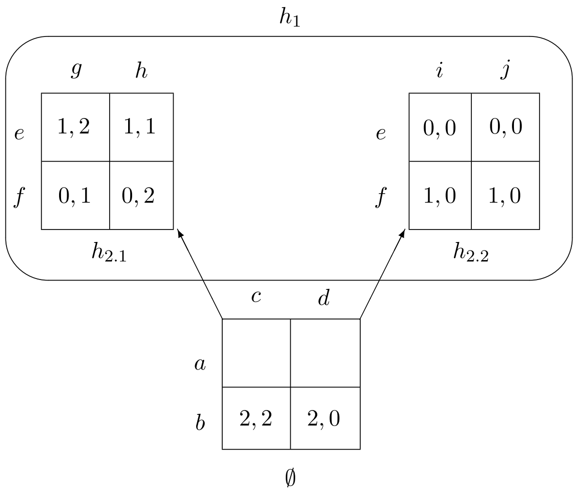


Fig. 1. Example of a dynamic game.

to player 1. Note that information sets $h_1, h_{2.1}$ and $h_{2.2}$ follow ∅, and that player 2's information sets $h_{2.1}$ and $h_{2.2}$ are simultaneous with player 1's information set h_1. At $h_1, h_{2.1}$ and $h_{2.2}$, players 1 and 2 simultaneously make a choice, after which the game ends.

2.2. Strategies

A strategy for player i is a complete choice plan, prescribing a choice at each of his information sets that can possibly be reached by this choice plan. Formally, for every $h_i, h'_i \in H_i$ such that h_i precedes h'_i, let $c_i(h_i, h'_i)$ be the choice at h_i for player i that leads to h'_i. Note that $c_i(h_i, h'_i)$ is unique by perfect recall. Consider a subset $\hat{H}_i \subseteq H_i$, not necessarily containing all information sets for player i, and a function s_i that assigns to every $h_i \in \hat{H}_i$ some choice $s_i(h_i) \in C_i(h_i)$. We say that s_i *allows* an information set $h_i \in H_i$ if at every $h'_i \in \hat{H}_i$ preceding h_i we have that $s_i(h'_i) = c_i(h'_i, h_i)$. By $H_i(s_i)$ we denote the collection of player i information sets that s_i allows for. A *strategy* for player i is a function s_i, assigning to every $h_i \in \hat{H}_i \subseteq H_i$ some choice $s_i(h_i) \in C_i(h_i)$, such that $\hat{H}_i = H_i(s_i)$. By S_i we denote the set of strategies for player i.

Note that this definition slightly differs from the standard definition of a strategy in the literature. Usually, a strategy for player i is defined as a mapping that assigns to *every* information set $h_i \in H_i$ some available choice – also to those information sets h_i that cannot be reached by s_i. The definition of a strategy we use corresponds to what Rubinstein (1991) calls a *plan of action*. One can also interpret it as the equivalence class of strategies (in the classical sense) that are outcome-equivalent. Hence, taking for every player the set of strategies as we use it corresponds to considering the pure strategy reduced normal form. However, for the concepts and results in this paper it does not matter which notion of strategy we use.

The following definitions will be especially important for what we do in this paper.

Definition 2.1 *(Strategies that allow an information set).* Consider an information set h. Then, we define the following sets:

$$S(h) = \left\{(s_i)_{i \in I} \in \prod_{i \in I} S_i\colon\ (s_i)_{i \in I} \text{ reaches some history in } h\right\},$$

$$S_i(h) = \left\{s_i \in S_i\colon\ (s_i, s_{-i}) \in S(h) \text{ for some } s_{-i} \in S_{-i}\right\},$$

$$S_{-i}(h) = \left\{s_{-i} \in S_{-i}\colon\ (s_i, s_{-i}) \in S(h) \text{ for some } s_i \in S_i\right\}.$$

Here, by S_{-i} we denote the set $\prod_{j \neq i} S_j$ of opponents' strategy combinations. We say that strategies in $S_i(h)$, and strategy combinations in $S_{-i}(h)$, *allow the information set* h. By perfect recall we have that $S(h_i) = S_i(h_i) \times S_{-i}(h_i)$ for every player i and every information set $h_i \in H_i$.

In the game of Fig. 1, the strategies for player 1 are $(a, e), (a, f)$ and b, whereas the strategies for player 2 are $(c, g), (c, h), (d, i)$ and (d, j). Note that within our terminology, b is a complete strategy for player 1 as player 1, by choosing b, will make sure that his subsequent information set h_1 cannot be reached, and hence we do not have to specify what player 1 would do if h_1 would be reached. Note also that player 1 cannot make his choice dependent on whether $h_{2.1}$ or $h_{2.2}$ is reached, since these are information sets for player 2 only, and player 1 does not know which of these information sets is reached. As such, (a, e) is a complete strategy for player 1. For player 2, (c, g) is a complete strategy as by choosing c player 2 will make sure that $h_{2.2}$ cannot be reached, and hence we do not have to specify what player 2 would do if $h_{2.2}$ would be reached. Similarly for his other three strategies.

Table 1
An epistemic model for the game in Fig. 1.

Types	$T_1 = \{t_1, t_1'\}$, $T_2 = \{t_2\}$
Beliefs for player 1	$b_1(t_1, \emptyset) = ((c, h), t_2)$ $b_1(t_1, h_1) = ((c, h), t_2)$ $b_1(t_1', \emptyset) = ((d, i), t_2)$ $b_1(t_1', h_1) = ((d, i), t_2)$
Beliefs for player 2	$b_2(t_2, \emptyset) = (b, t_1)$ $b_2(t_2, h_{2.1}) = ((a, f), t_1')$ $b_2(t_2, h_{2.2}) = ((a, f), t_1')$

In this example, the sets of strategies that allow the various information sets are as follows:

$$S_1(\emptyset) = S_1, \qquad S_2(\emptyset) = S_2,$$
$$S_1(h_1) = S_1(h_{2.1}) = S_1(h_{2.2}) = \big\{(a, e), (a, f)\big\},$$
$$S_2(h_1) = S_2, \qquad S_2(h_{2.1}) = \big\{(c, g), (c, h)\big\}, \qquad S_2(h_{2.2}) = \big\{(d, i), (d, j)\big\}.$$

2.3. Epistemic model

We now wish to model the players' beliefs in the game. At every information set $h_i \in H_i$, player i holds a belief about (a) the opponents' strategy choices, (b) the beliefs that the opponents have, at their information sets, about the other players' strategy choices, (c) the beliefs that the opponents have, at their information sets, about the beliefs their opponents have, at their information sets, about the other players' strategy choices, and so on. A possible way to represent such conditional belief hierarchies is as follows.

Definition 2.2 *(Epistemic model).* Consider a dynamic game Γ. An epistemic model for Γ is a tuple $M = (T_i, b_i)_{i \in I}$ where

(a) T_i is the finite set of types for player i,
(b) b_i is a function that assigns to every type $t_i \in T_i$, and every information set $h_i \in H_i$, a probability distribution $b_i(t_i, h_i) \in \Delta(S_{-i}(h_i) \times T_{-i})$.

Recall that $S_{-i}(h_i)$ represents the set of opponents' strategy combinations that allow h_i. By $T_{-i} := \prod_{j \neq i} T_j$ we denote the set of opponents' type combinations. For every finite set X, we denote by $\Delta(X)$ the set of probability distributions on X.

So, at every information set $h_i \in H_i$ type t_i holds a conditional probabilistic belief $b_i(t_i, h_i)$ about the opponents' strategies and types. In particular, type t_i holds conditional beliefs about the opponents' strategies. As every opponent's type holds conditional beliefs about the other players' strategies, every type t_i holds at every $h_i \in H_i$ also a conditional belief about the opponents' conditional beliefs about the other players' strategy choices. And so on. Since a type may hold different beliefs at different histories, a type may, during the game, revise his belief about the opponents' strategies, but also about the opponents' conditional beliefs.

For a given type t_i within an epistemic model, we can *derive* the complete belief hierarchy it induces. As an illustration, consider the epistemic model in Table 1, which is an epistemic model for the game in Fig. 1. So, we consider two possible types for player 1, t_1 and t_1', and one possible type for player 2, t_2. Player 2's type t_2 believes at the beginning of the game that player 1 chooses b and is of type t_1, whereas at $h_{2.1}$ and $h_{2.2}$ this type believes that player 1 chooses strategy (a, f) and is of type t_1'. In particular, type t_2 revises his belief about player 1's strategy choice if the game moves from $\emptyset$ to $h_{2.1}$ or $h_{2.2}$. Note that player 1's type t_1 believes that player 2 chooses strategy (c, h), whereas his other type t_1' believes that player 2 chooses strategy (d, i). So, type t_2 believes at $\emptyset$ that player 1 believes that player 2 chooses (c, h), whereas t_2 believes at $h_{2.1}$ and $h_{2.2}$ that player 1 believes that player 2 chooses (d, i). Hence, player 2's type t_2 also revises his belief about player 1's *belief* if the game moves from $\emptyset$ to $h_{2.1}$ or $h_{2.2}$. By continuing in this fashion, we can derive the full belief hierarchy for type t_2. Similarly for the other types in this model.

Note that in our definition of an epistemic model, we restrict attention to models with *finitely many types*. The reason is that it suffices for the things we want to do in this paper.

3. Belief in the opponents' future rationality

We now present the main idea in this paper, namely that a player always believes that his opponents will choose rationally now and in the future. We first define what it means for a strategy s_i to be optimal for a type t_i at a given information set h. Consider a type t_i, a strategy s_i and an information set $h_i \in H_i(s_i)$ that is allowed by s_i. By $u_i(s_i, t_i \mid h_i)$ we denote the expected utility from choosing s_i under the conditional belief that t_i holds at h_i about the opponents' strategy choices.

Definition 3.1 *(Optimality at a given information set).* Consider a type t_i, a strategy s_i and a history $h_i \in H_i(s_i)$. Strategy s_i is optimal for type t_i at h_i if $u_i(s_i, t_i \mid h_i) \geqslant u_i(s_i', t_i \mid h_i)$ for all $s_{\in}' \in S_i(h_i)$.

Remember that $S_i(h_i)$ is the set of player i strategies that allow h_i. We can now define belief in the opponents' future rationality.

Definition 3.2 *(Belief in the opponents' future rationality).* Consider a type t_i, an information set $h_i \in H_i$, and an opponent $j \neq i$. Type t_i believes at h_i in j's future rationality if $b_i(t_i, h_i)$ only assigns positive probability to j's strategy-type pairs (s_j, t_j) where s_j is optimal for t_j at every $h_j' \in H_j(s_j)$ that weakly follows h_i. Type t_i believes in the opponents' future rationality if at every $h_i \in H_i$, type t_i believes in every opponent's future rationality.

So, to be precise, a type that believes in the opponents' future rationality believes that every opponent chooses rationally now (if the opponent makes a choice at a simultaneous information set), and at every information set that follows. As such, the correct terminology would be "belief in the opponents' *present* and future rationality", but we stick to "belief in the opponents' future rationality" as to keep the name short.

In the epistemic model of Table 1, it may be verified that the type t_1 for player 1, and the type t_2 for player 2, believe in the opponent's future rationality. For instance, type t_1 believes at $\emptyset$ and h_1 that player 2 chooses strategy (c, h) and is of type t_2. On the other hand, type t_2 believes at $\emptyset$ that player 1 chooses b, and believes at $h_{2.1}$ that player 1 chooses (a, f). So, strategy (c, h) is optimal for t_2 at $\emptyset$, and is also optimal for t_2 at $h_{2.1}$. As such, t_1 believes at $\emptyset$ and h_1 in player 2's future rationality. Similarly, it can be verified that player 2's type t_2 believes in 1's future rationality.

However, player 1's type t_1' does not believe in 2's future rationality, as type t_1' believes at $\emptyset$ that player 2 chooses strategy (d, i) and is of type t_2, whereas strategy (d, i) is clearly not optimal for t_2 at $\emptyset$.

Another condition we impose is that a type uses Bayesian updating when revising his belief about the opponent's strategy and type, whenever this is possible. It is well-known that with this requirement, a player can always construct a strategy that is optimal at *each* of his information sets – something that is not generally possible if we do not impose Bayesian updating. Moreover, under Bayesian updating such an optimal strategy can be designed by recursively selecting optimal *choices* at all information sets, starting at the final information sets of this player and then working backwards until his first information sets are reached. This result is often referred to as the *one-deviation principle* (see Hendon et al., 1996 and Perea, 2002). Formally, Bayesian updating is defined as follows.

Definition 3.3 *(Bayesian updating).* A type t_i satisfies Bayesian updating if for every two informations sets $h_i, h_i' \in H_i$ where h_i' follows h_i, and $b_i(t_i, h_i)(S_{-i}(h_i') \times T_{-i}) > 0$, it holds that

$$b_i\big(t_i, h_i'\big)(s_{-i}, t_{-i}) = \frac{b_i(t_i, h_i)(s_{-i}, t_{-i})}{b_i(t_i, h_i)(S_{-i}(h_i') \times T_{-i})}$$

for every opponents' strategy–type combination $(s_{-i}, t_{-i}) \in S_{-i}(h_i') \times T_{-i}$.

Here, $S_{-i}(h_i')$ denotes the set of opponents' strategy combinations that allow h_i', and T_{-i} denotes the set of opponents' type combinations. By $b_i(t_i, h_i)(S_{-i}(h_i') \times T_{-i})$ we denote the total probability that the conditional belief $b_i(t_i, h_i)$ assigns to all opponents' strategy–type combinations in $S_{-i}(h') \times T_{-i}$.

Next, we formalize the requirement that a player not only believes in the opponents' future rationality and satisfies Bayes updating, but also always believes that every opponent does so, and so on. This leads to the definition of *common belief in future rationality*. We proceed by recursively defining sets of types T_i^k for all $k \in \mathbb{N}$.

Definition 3.4 *(Common belief in future rationality).* Consider a dynamic game Γ and some epistemic model $M = (T_i, b_i)_{i \in I}$ for Γ.

Initial step. Define for every player i the set of types

$$T_i^1 := \{t_i \in T_i\colon\ t_i \text{ believes in the opponents' future rationality and satisfies Bayesian updating}\}.$$

Inductive step. Let $k \geqslant 2$, and suppose that T_i^{k-1} has been defined for all players i. Then, we define

$$T_i^k := \big\{t_i \in T_i^{k-1}\colon\ b_i(t_i, h_i)\big(S_{-i} \times T_{-i}^{k-1}\big) = 1 \text{ for all } h_i \in H_i\big\}.$$

A type t_i expresses common belief in future rationality if $t_i \in T_i^k$ for every k.

Here, $T_{-i}^{k-1} := \prod_{j \neq i} T_j^{k-1}$. So, a type in T_i^2 always believes that his opponents' believe in their opponents' future rationality and that his opponents satisfy Bayesian updating. A type in T_i^3 always believes, in addition, that his opponents reason in this way as well, and so on. We say that types in T_i^k express *k-fold belief in future rationality*.

In the epistemic model of Table 1, no type expresses common belief in future rationality. Namely, type t_1' does not believe in 2's future rationality as we have seen. As player 2's type t_2 believes at $h_{2.1}$ and $h_{2.2}$ that player 1 is of type t_1', type t_2 does not express 2-fold belief in future rationality. Type t_1 believes that player 2 is of type t_2, and hence does not express 3-fold belief in future rationality. We may thus conclude that no type in that model expresses common belief in future rationality.

Finally, we define those strategies that can rationally be chosen under common belief in future rationality. Before doing so, we first state what it means for a strategy to be rational for a type.

Definition 3.5 *(Rational strategy).* A strategy s_i is rational for a type t_i if s_i is optimal for t_i at every $h_i \in H_i(s_i)$.

In the literature, this is often called *sequential rationality*. A strategy should thus be optimal at every information set that is allowed by this strategy, given the conditional belief that is held at that information set.

Definition 3.6 *(Rational strategy under common belief in future rationality).* A strategy s_i can rationally be chosen under common belief in future rationality if there is some epistemic model $M = (T_i, b_i)_{i \in I}$, and some type $t_i \in T_i$, such that t_i expresses common belief in future rationality, and s_i is rational for t_i.

In other words, a strategy can rationally be chosen under common belief in future rationality if there is some type expressing common belief in future rationality that supports this strategy choice.

4. Penta's backwards rationalizability procedure

We will now provide a recursive procedure that yields precisely those strategies – and conditional beliefs about the opponents' strategies – that are possible under *common belief in future rationality*. The procedure is due to Penta (2009) who calls it the *backwards rationalizability* procedure.

4.1. Description of the procedure

To formally define the backwards rationalizability procedure we first introduce the notion of a *conditional belief vector*, and define what it means for a conditional belief vector to satisfy Bayesian updating.

Definition 4.1 *(Conditional belief vector).* A conditional belief vector for player i is a vector $b_i = (b_i(h_i))_{h_i \in H_i}$ that assigns to every information set $h_i \in H_i$ a probabilistic belief $b_i(h_i) \in \Delta(S_{-i}(h_i))$ about the opponents' strategy combinations that allow h_i.

The conditional belief vector $b_i = (b_i(h_i))_{h_i \in H_i}$ satisfies Bayesian updating if for every two informations sets $h_i, h_i' \in H_i$ where h_i' follows h_i, and $b_i(h_i)(S_{-i}(h_i')) > 0$, it holds that

$$b_i(h_i')(s_{-i}) = \frac{b_i(h_i)(s_{-i})}{b_i(h_i)(S_{-i}(h_i'))}$$

for every opponents' strategy combination $s_{-i} \in S_{-i}(h_i')$.

Let us denote by B_i the set of all conditional belief vectors for player i in the game Γ. Penta's recursive procedure eliminates at every step some strategies *and conditional belief vectors* from the game, and is formally defined as follows.

Algorithm 4.2 *(Penta's backwards rationalizability procedure).*

Initial step. *For every player i and all information sets $h \in H$ define*

$$S_i^0(h) := S_i(h),$$
$$B_i^0 := \{b_i \in B_i\colon\ b_i \text{ satisfies Bayesian updating}\}.$$

Inductive step. *Let $k \geqslant 1$, and suppose that $S_i^{k-1}(h)$ and B_i^{k-1} have been defined for all players i, and all $h \in H$. Then, we define for all players i and all information sets $h \in H$*

$$S_i^k(h) := \{s_i \in S_i^{k-1}(h)\colon\ \text{there is some } b_i \in B_i^{k-1} \text{ such that } s_i \text{ is optimal for } b_i(h_i') \text{ at every } h_i' \succcurlyeq h \text{ with } h_i' \in H_i(s_i)\},$$
$$B_i^k := \{b_i \in B_i^{k-1}\colon\ b_i(h_i) \in \Delta(S_{-i}^k(h_i)) \text{ for all } h_i \in H_i\}.$$

A strategy s_i is said to survive the backwards rationalizability procedure if $s_i \in S_i^k(\emptyset)$ for every k.

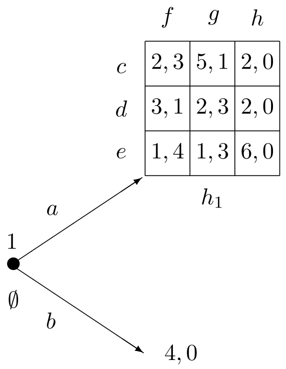


Fig. 2. Illustration of the two procedures.

Here, we denote by H the collection of all information sets in the game. Remember that $h_i' \succcurlyeq h$ means that h_i' weakly follows h. So, in the backwards rationalizability procedure we need to keep track of the players' possible *conditional belief vectors* at every stage of the process, and not only of the possible strategies at the various information sets. Since the game Γ is finite, it may easily be verified that this procedure will stop after finitely many rounds.

Penta (2009) uses this procedure also for games with incomplete information, and applies it to problems of mechanism design and implementation. Hence, he uses it for a much wider range of situations than we do here in this paper. Moreover, Penta shows that the backwards rationalizability procedure characterizes the epistemic concept of *interim perfect equilibrium* across models of beliefs.

4.2. Illustration of the procedure

We will now illustrate the backwards rationalizability procedure by means of an example. Consider the game in Fig. 2. So, at the beginning of the game, $\emptyset$, only player 1 is active. He can choose between a and b. If he chooses b, the game ends and the utilities are 4 and 0 for the players. If he chooses a, then we reach information set h_1 at which players 1 and 2 choose simultaneously. At h_1, player 1 is the row player, and player 2 the column player.

There are two information sets in this game, $\emptyset$ and h_1. Clearly, strategy (a,d) cannot be optimal for player 1 at $\emptyset$ for any conditional belief, whereas h cannot be optimal for player 2 at h_1 for any conditional belief. However, (a,d) *can* be optimal for player 1 at h_1 for some conditional belief. So, we have that

$$\begin{aligned}
&S_1^1(\emptyset) = \big\{(a,c),(a,e),b\big\},\\
&S_1^1(h_1) = \big\{(a,c),(a,d),(a,e)\big\},\\
&S_2^1(\emptyset) = \{f,g\},\\
&S_2^1(h_1) = \{f,g\}.
\end{aligned}$$

As a consequence,

$$\begin{aligned}
&B_1^1 = \big\{b_1 \in B_1^0\colon\ b_1(\emptyset) \in \Delta\big(\{f,g\}\big) \text{ and } b_1(h_1) \in \Delta\big(\{f,g\}\big)\big\},\\
&B_2^1 = B_2^0.
\end{aligned}$$

After this step, strategy (a,e) cannot be optimal at $\emptyset$ – nor at h_1 – for any conditional belief vector in B_1^1. So, we have that

$$\begin{aligned}
&S_1^2(\emptyset) = \big\{(a,c),b\big\},\\
&S_1^2(h_1) = \big\{(a,c),(a,d)\big\},\\
&S_2^2(\emptyset) = \{f,g\},\\
&S_2^2(h_1) = \{f,g\}.
\end{aligned}$$

As a consequence,

$$\begin{aligned}
&B_1^2 = B_1^1,\\
&B_2^2 = \big\{b_2 \in B_2^0\colon\ b_2(h_1) \in \Delta\big(\big\{(a,c),(a,d)\big\}\big)\big\}.
\end{aligned}$$

After this step the procedure stops, as no further strategies or beliefs can be eliminated. We thus conclude that the strategies which survive the backwards rationalizability procedure are (a,c) and b for player 1, and f and g for player 2. It may be verified that these are precisely the strategies that players 1 and 2 can rationally choose under common belief in future rationality.

Note that the concept of *sequential equilibrium* singles out the strategy b for player 1. Namely, in the subgame at h_1 the only Nash equilibrium is $(\frac{1}{2}c + \frac{1}{2}d, \frac{3}{4}f + \frac{1}{4}g)$. Hence, in a sequential equilibrium, player 1 must believe that, with probability $\frac{3}{4}$, player 2 will choose f and with probability $\frac{1}{4}$ he will choose g. As such, player 1's expected utility from choosing a at the beginning will be $11/4 < 4$, and therefore player 1 must choose b.

But why should player 1 exactly attribute the probabilities $\frac{3}{4}$ and $\frac{1}{4}$ to the strategies f and g? The fact that player 1 may assign a positive probability to g indicates that apparently g is a reasonable choice for player 2. But why could player 1 then not assign probability 1 to g, and choose (a, c) as a best response to that?

Common belief in future rationality allows player 1 to choose strategy (a, c), because under this concept he may indeed believe that player 2 will choose g with probability 1. So, in this example sequential equilibrium is really more restrictive than common belief in future rationality. In fact, we believe that sequential equilibrium is *too* restrictive in this example.

4.3. Characterization result

The following theorem shows that the backwards rationalizability procedure yields precisely those strategies that can rationally be chosen under common belief in future rationality.

Theorem 4.3 *(Strategies selected by backwards rationalizability procedure). A strategy s_i can rationally be chosen under common belief in future rationality, if and only if, s_i survives the backwards rationalizability procedure.*

The proof is very similar to the proof of Theorem 5.4 which we will present in the following section, and is therefore omitted.

5. Backward dominance procedure

We will now provide a recursive procedure, called the *backward dominance procedure*, that proceeds by *eliminating strategies only* – not beliefs – and where strategies are removed based on strict dominance arguments. In that sense, the algorithm is similar to the *iterated conditional dominance procedure* by Shimoji and Watson (1998) which characterizes those strategies that are possible under *extensive-form rationalizability* (Pearce, 1984; Battigalli, 1997). Since Battigalli and Siniscalchi (2002) prove that the strategies that can rationally be chosen under the epistemic concept of *common strong belief in rationality* are precisely the extensive-form rationalizable strategies, the iterated conditional dominance procedure also selects precisely those strategies that players can rationally choose under *common strong belief in rationality*.

Later on in this section we will prove that every strategy that can rationally be chosen under common belief in future rationality will always survive the backward dominance procedure, but not necessarily *vice versa*. So, the backward dominance procedure always gives a superset of the set of strategies that are possible under common belief in future rationality. Consequently, we can use the backward dominance procedure as a *first selection procedure* to find those strategies that players can choose under common belief in future rationality.

In fact, the only difference between the backward dominance procedure and common belief in future rationality is that the former allows for conditional beliefs that violate Bayesian updating, whereas the latter does not. More precisely, we will prove that the backward dominance procedure yields *precisely* those strategies that are possible if we would impose common belief in future rationality, but without the Bayesian updating requirement. In many games, this difference has no effect on the possible strategies that players can choose under common belief in future rationality, but there are games where this difference is important – as we will see.

5.1. Description of the procedure

In order to formally state our algorithm we need the following definitions. For a given information set h let $S(h)$ be the set of all strategy combinations $s = (s_i)_{i \in I}$ that reach h. Then, a subset $\Gamma(h) \subseteq S(h)$ is called a *decision problem at* h if for every active player i at h there are some sets $D_i \subseteq S_i(h)$ and $D_{-i} \subseteq S_{-i}(h)$ such that $\Gamma(h) = D_i \times D_{-i}$. The intuition is that player i at h believes that his opponents will choose some strategy combination in D_{-i}, whereas player i himself only considers strategies in D_i as reasonable options. So, in a sense, D_i represents the objects of choice for player i at h, whereas D_{-i} represents the states of the world about which he is uncertain. This explains the term *decision problem*.

Consider an information set h, a player i who is active at h, and a decision problem $\Gamma(h) = D_i \times D_{-i}$ at h. We say that strategy $s_i \in D_i$ is *strictly dominated* within the decision problem $\Gamma(h)$ if there is some randomized strategy $\mu_i \in \Delta(D_i)$ such that $u_i(\mu_i, s_{-i}) > u_i(s_i, s_{-i})$ for all $s_{-i} \in D_{-i}$. By $sd_i(\Gamma(h))$ we denote the set of strategies $s_i \in D_i$ that are strictly dominated within $\Gamma(h)$ for the active player i. Similarly, we denote by

$$sd\big(\Gamma(h)\big) := \big\{(s_i)_{i \in I} \in \Gamma(h)\colon\ s_i \in sd_i\big(\Gamma(h)\big) \text{ for some } i \text{ that is active at } h\big\}$$

the set of strategy combinations in $\Gamma(h)$ that involve a strategy for an active player that is strictly dominated within $\Gamma(h)$.

Remember that $h' \succcurlyeq h$ means that information set h' weakly follows information set h. The *backward dominance procedure* can now be defined as follows.

Algorithm 5.1 *(Backward dominance procedure).*

Initial step. *For every information set h, define* $\Gamma^0(h) := S(h)$.

Inductive step. *Let* $k \geqslant 1$*, and suppose that the decision problems* $\Gamma^{k-1}(h)$ *have been defined for every information set h. Then, at every information set h we define*

$$\Gamma^k(h) := \Gamma^{k-1}(h) \backslash \bigcup_{h' \succcurlyeq h} sd\big(\Gamma^{k-1}(h')\big).$$

A strategy s_i *survives the backward dominance procedure if there is some* $s_{-i} \in S_{-i}$ *such that* $(s_i, s_{-i}) \in \Gamma^k(\emptyset)$ *for all k.*

So, at every step of the procedure we delete from the current decision problem $\Gamma^{k-1}(h)$ those strategies s_i that are strictly dominated within a decision problem $\Gamma^{k-1}(h')$ at which player i is active, and that weakly follows h. Since we only have finitely many strategies in the game, and the decision problems can only become smaller at every step, this procedure must converge after finitely many steps.

However, in order for this procedure to be well-defined, we must show that $\Gamma^k(h)$ is a *decision problem* for every k and every information set h. That is, we must show that, whenever i is an active player at h, then $\Gamma^k(h) = D_i \times D_{-i}$ for some $D_i \subseteq S_i(h)$ and $D_{-i} \subseteq S_{-i}(h)$. The following lemma guarantees that this is always true.

Lemma 5.2 *(Every* $\Gamma^k(h)$ *is a decision problem). For every* $k \geqslant 0$*, every information set h, and every active player i at h, there are sets* $D_i \subseteq S_i(h)$ *and* $D_{-i} \subseteq S_{-i}(h)$ *such that* $\Gamma^k(h) = D_i \times D_{-i}$.

The proof can be found in Section 9. Another important question is whether this procedure always delivers a *nonempty* set of strategies for every player at every information set. Or is it possible that at a given information set we delete all strategies for a player? We will see that the algorithm will never eliminate all strategies for a player at an information set. Here, we say that a strategy s_i *survives* the backward dominance procedure at some information set h if there is some $s_{-i} \in S_{-i}$ such that $(s_i, s_{-i}) \in \Gamma^k(h)$ for all k.

Theorem 5.3 *(Backward dominance delivers nonempty output). For every information set h, and every player i, there is at least one strategy* $s_i \in S_i(h)$ *that survives the backward dominance procedure at h.*

The formal proof for this result can be found in Section 9.

5.2. Illustration of the procedure

We will now illustrate our backward dominance procedure by means of the game in Fig. 2. At the beginning of the procedure we start with two decision problems, namely the full decision problem $\Gamma^0(\emptyset)$ at $\emptyset$ where only player 1 is active, and the full decision problem $\Gamma^0(h_1)$ at h_1 where both players are active. These decision problems can be found in Table 2.

Step 1. At $\Gamma^0(\emptyset)$ we delete strategy (a, d) for player 1 since it is strictly dominated at $\Gamma^0(\emptyset)$ by b. At $\Gamma^0(\emptyset)$ we also delete strategy h for player 2 since it is strictly dominated by f and g at the future decision problem $\Gamma^0(h_1)$ at which player 2 is active. Finally, at $\Gamma^0(h_1)$ we delete strategy h for player 2 as it is strictly dominated by f and g at $\Gamma^0(h_1)$. This leads to the new decision problems $\Gamma^1(\emptyset)$ and $\Gamma^1(h_1)$ which can be found in Table 3.

Step 2. At $\Gamma^1(\emptyset)$ we delete strategy (a, e) for player 1 since it is strictly dominated at $\Gamma^1(\emptyset)$ by (a, c) and b. At $\Gamma^1(h_1)$ we delete strategy (a, e) for player 1 since it is strictly dominated by (a, c) and (a, d) at $\Gamma^1(h_1)$. This leads to the new decision problems $\Gamma^2(\emptyset)$ and $\Gamma^2(h_1)$ presented in Table 4.

After this step no more strategies can be eliminated. So, the algorithm stops here, and the strategies that survive the backward dominance procedure are (a, c) and b for player 1, and f and g for player 2. Note that these are the same strategies that survived Penta's backwards rationalizability procedure.

Table 2
Full decision problems in backward dominance procedure:

Player 1 active

$\Gamma^0(\emptyset)$	f	g	h
(a, c)	2, 3	5, 1	2, 0
(a, d)	3, 1	2, 3	2, 0
(a, e)	1, 4	1, 3	6, 0
b	4, 0	4, 0	4, 0

Players 1 and 2 active

$\Gamma^0(h_1)$	f	g	h
(a, c)	2, 3	5, 1	2, 0
(a, d)	3, 1	2, 3	2, 0
(a, e)	1, 4	1, 3	6, 0

Table 3
Step 1 of backward dominance procedure:

Player 1 active

$\Gamma^1(\emptyset)$	f	g
(a,c)	$2,3$	$5,1$
(a,e)	$1,4$	$1,3$
b	$4,0$	$4,0$

Players 1 and 2 active

$\Gamma^1(h_1)$	f	g
(a,c)	$2,3$	$5,1$
(a,d)	$3,1$	$2,3$
(a,e)	$1,4$	$1,3$

Table 4
Step 2 of backward dominance procedure:

Player 1 active

$\Gamma^2(\emptyset)$	f	g
(a,c)	$2,3$	$5,1$
b	$4,0$	$4,0$

Players 1 and 2 active

$\Gamma^2(h_1)$	f	g
(a,c)	$2,3$	$5,1$
(a,d)	$3,1$	$2,3$

5.3. Characterization result

It turns out that the backward dominance procedure eventually yields a set of strategies for every player that *includes* the strategies that can rationally be chosen under *common belief in future rationality*. So, the backward dominance procedure can be used as a first selection procedure to eventually find all strategies that are possible under common belief in future rationality. This is an important result, since the backward dominance procedure is very easy to use – the only thing we have to do is to recursively eliminate strategies at information sets based on strict dominance criteria. In fact, the whole procedure can be formulated as a linear program, and hence can be implemented easily on a computer. In contrast, the backwards rationalizability procedure requires us to simultaneously remove strategies *and conditional belief vectors* at every step, and does not work with strict dominance arguments, which makes it more difficult to use than the backward dominance procedure.

We can actually be more precise about the strategies selected by the backward dominance procedure – they turn out to be precisely those strategies that a player can rationally choose if we would impose common belief in future rationality *without the Bayesian updating requirement*. So, what separates the backward dominance procedure from common belief in future rationality is (common belief in) Bayesian updating – nothing more and nothing less. Consequently, if we want to find precisely those strategies that players can rationally choose under common belief in future rationality (with the Bayesian updating condition), then we could first run the backward dominance procedure, and afterwards additionally impose common belief in Bayesian updating.

Theorem 5.4 *(Strategies selected by backward dominance procedure). Player i can rationally choose strategy s_i under common belief in future rationality without the Bayesian updating requirement, if and only if, s_i survives the backward dominance procedure.*

The proof can be found in Section 9.

6. Discussion

In this section we will discuss some important properties of the concept of common belief in future rationality, and of the associated backward dominance procedure and backwards rationalizability procedure.

6.1. Bayesian updating can matter

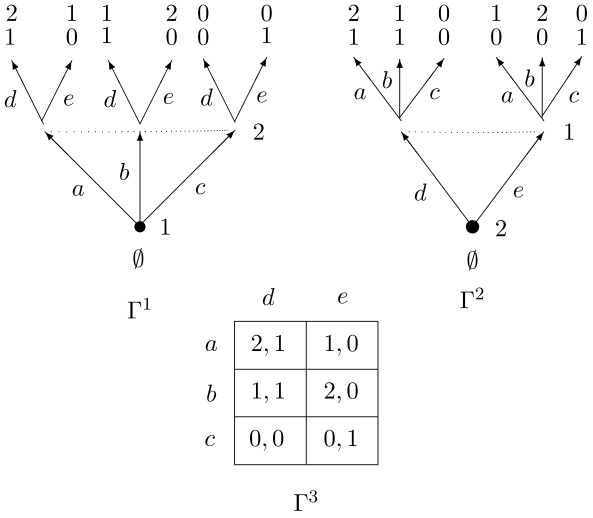
Shimoji and Watson (1998) have shown that for the concept of *extensive-form rationalizability* (Pearce, 1984; Battigalli, 1997) it is inessential whether we impose (common belief in) Bayesian updating or not. As the epistemic notion of *common strong belief in rationality* (Battigalli and Siniscalchi, 2002) selects precisely the extensive-form rationalizable strategies, we may conclude that Bayesian updating is also irrelevant for the eventual strategy choices selected by *common strong belief in rationality* (although it may matter for the *types* selected by this concept).

This naturally raises the question whether the same is true for the concept of *common belief in future rationality*. So, would it matter for the eventual strategy choices of the players whether we drop the Bayesian updating condition or not? We will see that Bayesian updating *can* matter for the concept of common belief in future rationality.

Consider, for instance, the game in Fig. 1. We will show that player 2 can rationally choose (c,h) under common belief in future rationality if we would *not* impose (common belief in) Bayesian updating, but that he can longer rationally choose (c,h) if we do impose (common belief in) Bayesian updating. To that purpose, consider the epistemic model in Table 5. It may be verified that both types, t_1 and t_2, express common belief in future rationality without the Bayesian updating requirement. As strategy (c,h) is rational for type t_2, we may conclude that player 2 can rationally choose strategy (c,h) under common belief in future rationality without the Bayesian updating requirement. Note, however, that type t_1 does not satisfy Bayesian updating when the game moves from $\emptyset$ to h_1. So, type t_2 does not believe that player 1 satisfies Bayesian updating.

Table 5
A new epistemic model for the game in Fig. 1.

Types	$T_1 = \{t_1\}$, $T_2 = \{t_2\}$
Beliefs for player 1	$b_1(t_1, \emptyset) = ((c, h), t_2)$ $b_1(t_1, h_1) = ((d, i), t_2)$
Beliefs for player 2	$b_2(t_2, \emptyset) = (b, t_1)$ $b_2(t_2, h_{2.1}) = ((a, f), t_1)$ $b_2(t_2, h_{2.2}) = ((a, f), t_1)$



	d	e
a	2, 1	1, 0
b	1, 1	2, 0
c	0, 0	0, 1

Fig. 3. Chronological order of moves matters for "common belief in future rationality".

In fact, we can show that player 2 can no longer rationally choose (c, h) if we additionally impose common belief in Bayesian updating. Namely, if player 1 believes in 2's future rationality, then player 1 must at $\emptyset$ assign probability 1 to player 2 choosing c. If we assume that player 1 satisfies Bayesian updating, then he must at h_1 still believe that player 2 has chosen c. Hence, at h_1 player 1 must believe that he is at $h_{2.1}$. So, player 1's unique optimal choice at h_1 is e. But then, if player 2 believes at $h_{2.1}$ that (a) player 1 believes at $\emptyset$ in player 2's future rationality, (b) player 1 satisfies Bayesian updating and (c) player 1 chooses rationally at h_1 (which weakly follows $h_{2.1}$), then player 2 must believe at $h_{2.1}$ that player 1 chooses e. Hence, player 2's unique optimal choice at $h_{2.1}$ is g. As player 2 will clearly choose c at $\emptyset$, we thus see that under common belief in future rationality *with* the Bayesian updating requirement, player 2 can only rationally choose the strategy (c, g). In particular, strategy (c, h) is no longer rational for player 2 under common belief in future rationality with the Bayesian updating requirement.

Equivalently, it can be shown that the backwards rationalizability procedure uniquely selects the strategy (c, g) for player 2 in this game, whereas the backward dominance procedure selects strategies (c, g) and (c, h) for player 2. The reader may verify this.

6.2. Modeling the order of moves

The concept of common belief in future rationality is very sensitive to the way in which we model the chronological order of moves in the game! Consider, for instance, the three games in Fig. 3.

In game Γ^1 player 1 moves before player 2, in game Γ^2 player 2 moves before player 1, and in game Γ^3 both players choose simultaneously. In Γ^1 and Γ^2, the second mover does not know which choice has been made by the first mover. So, all three games represent a situation in which both players choose in complete ignorance of the opponent's choice. Since the utilities in the games are identical, one can argue that these three games are in some sense "equivalent". In fact, the three games above only differ by applying the transformation of *interchange of decision nodes*,[2] as defined by Thompson (1952). However, for the concept of *common belief in future rationality* it crucially matters which of the three representations Γ^1, Γ^2 or Γ^3 we choose.

In the game Γ^1, common belief in future rationality does not restrict player 2's belief at all, as player 1 moves before him. So, player 2 can rationally choose d and e under common belief in future rationality here. On the other hand, player 1

[2] For a formal description of this transformation, the reader may consult Thompson (1952), Elmes and Reny (1994) or Perea (2001).

may believe that player 2 chooses d or e under common belief in future rationality, and hence player 1 himself may rationally choose a or b under common belief in future rationality.

In the game Γ^2, common belief in future rationality does not restrict player 1's beliefs as he moves after player 2. Hence, player 1 may rationally choose a or b under common belief in future rationality. Player 2 must therefore believe that player 1 will either choose a or b in the future, and hence player 2 can only rationally choose d under common belief in future rationality.

In the game Γ^3, finally, player 1 can only rationally choose a, and player 2 can only rationally choose d under common belief in future rationality. Namely, if player 2 believes in player 1's (present and) future rationality, then player 2 believes that player 1 does not choose c, since player 1 moves at the same time as player 2. Therefore, player 2 can only rationally choose d under common belief in future rationality. If player 1 believes in player 2's (present and) future rationality, and believes that player 2 believes in player 1's (present and) future rationality, then player 1 believes that player 2 chooses d, and therefore player 1 can only rationally choose a under common belief in future rationality.

Hence, the precise order of moves is very important for the concept of common belief in future rationality! In particular, this concept is *not invariant* with respect to Thompson's (1952) transformation of *interchange of decision nodes*.

Now, why is this? In the concept of common belief in future rationality, the key condition is that a player, at any point in time, believes that his opponent will choose rationally *from now on*, but not necessarily that he has acted rationally in the past – even if believing so is possible. Hence, for this concept it is absolutely crucial how we model the chronological order of moves. Indeed, if we interchange some decision nodes – as we did in the games in Fig. 3 – then we change the meaning of past, present and future in the game, and thereby also change the meaning of *belief in future rationality*! So, for the concept of common belief in future rationality, the interchange of decision nodes *does* change some essential features of the game.

Is this a big problem? Not if we take the precise modeling of the dynamic game seriously. Namely, if we insist that the order of the information sets in the dynamic game faithfully represents the actual chronological order of moves, then there is no problem in using common belief in future rationality as a concept.

6.3. Order independence

As we defined it, the backward dominance algorithm eliminates, at every step and every information set h, *all* strategies for player i that are strictly dominated at some decision problem for player i weakly following h. Suppose we would, at every step, only eliminate *some* of these strategies, but not all. Would it matter for the eventual result? The answer is "no": The order and speed in which we eliminate strategies in the backward dominance procedure has no influence on the final output. Here is an argument.

Let us compare two procedures, Procedure 1 and Procedure 2, where Procedure 1 eliminates, at every step, *all* strategies that can possibly be eliminated, whereas Procedure 2 eliminates at every step only *some* strategies that can be eliminated. Then, Procedure 1 will, at every step and every information set h, have eliminated at least as many strategies as Procedure 2. Namely, at Step 1 this is true by construction. Consider now Step 2. Suppose that in Procedure 2 we would eliminate strategy s_i at h because it is strictly dominated at the future decision problem $\tilde{\Gamma}^1(h')$ for player i. Here, $\tilde{\Gamma}^1(h')$ is the decision problem at h' after Step 1 of Procedure 2. Now, let $\Gamma^1(h')$ be the decision problem at h' after Step 1 of Procedure 1. Then, $\Gamma^1(h')$ contains at most as many strategies for i's opponents as $\tilde{\Gamma}^1(h')$. Hence, if s_i was strictly dominated at $\tilde{\Gamma}^1(h')$, it will certainly be strictly dominated at $\Gamma^1(h')$, and so in Procedure 1 we will also eliminate strategy s_i at h. We thus see that in Step 2, every strategy that is eliminated in Procedure 2 will also be eliminated in Procedure 1. Of course we can iterate this argument and conclude that at every step, Procedure 1 will have deleted as least as many strategies as Procedure 2.

We now show that the converse is also true, namely every strategy that is eliminated in Procedure 1 will also *eventually* be eliminated in Procedure 2. Suppose this would not be true. Then, let k be the last step such that every strategy eliminated by Procedure 1 *before* Step k is also eventually eliminated by Procedure 2. Take then a strategy s_i that is eliminated at some information set h in Step k of Procedure 1, but which is never eliminated in Procedure 2. The reason for eliminating s_i at h in Procedure 1 is that s_i is strictly dominated at some decision problem $\Gamma^{k-1}(h')$ for player i weakly following h. By assumption, in Procedure 2 there is some step $m \geqslant k-1$ such that the associated decision problem $\tilde{\Gamma}^m(h')$ is a "subset" of $\Gamma^{k-1}(h')$, which means that the strategy sets in $\tilde{\Gamma}^m(h')$ are contained in the strategy sets of $\Gamma^{k-1}(h')$. But then, if s_i is strictly dominated at $\Gamma^{k-1}(h')$, it is certainly strictly dominated in $\tilde{\Gamma}^m(h')$. As such, Procedure 2 must eliminate s_i sooner or later at information set h. This contradicts our assumption above. We may thus conclude that every strategy that is eliminated in Procedure 1 will also eventually be eliminated in Procedure 2.

Altogether, we see that Procedure 1 and Procedure 2 must eventually yield the same set of strategies at every information set. So, the order and speed in which we delete strategies from the game does not matter for the output of the backward dominance procedure. The intuitive reason is that the algorithm is *monotonic* in the following sense: If we make the decision problems smaller, then it becomes easier for a strategy to become strictly dominated, and hence we will eliminate more, which in turn leads to smaller decision problems, and so on.

In fact, the same conclusion can be drawn for the backwards rationalizability procedure, based on a similar reasoning. Also for that procedure, the order and speed in which we delete strategies and conditional belief vectors from the game is irrelevant.

This result also has some important *practical* implications for the backward dominance procedure and the backwards rationalizability procedure. In some games it may be easier not to eliminate strategies at *all* information sets simultaneously, but rather to start with the decision problems at the end of the game, apply the procedure there until we can eliminate nothing more, then turn to decision problems that come just before, apply the procedure there until we can eliminate nothing more, and so on. That is, to use a *backward induction approach* to eliminate the strategies. Such an order of elimination will be convenient especially for large dynamic games, with many consecutive information sets. In fact, Penta (2009) refers to this backward induction approach as the "backwards procedure" if we apply it to the backwards rationalizability procedure.

6.4. *Games with perfect information*

In this section we explore what common belief in future rationality does for games with *perfect information*. A dynamic game is said to be with *perfect information* if at every information set exactly one player is active, and this player knows exactly which choices have been made until then. Formally, this means that at every information set h there is exactly one player i with $h \in H_i$, and the information set h consists of a single history x.

As in Battigalli (1997), we say that a game with perfect information is *with no relevant ties* if for every player i, and every information set $h_i \in H_i$, two different choices at h_i will always lead to two different utilities for player i. That is, for every two terminal histories z, z' following $h_i \in H_i$ which contain different choices at h_i, we have that $u_i(z) \neq u_i(z')$. It is easily seen that every game with perfect information and no relevant ties yields a unique backward induction strategy for every player.

Consider now an arbitrary game with perfect information and no relevant ties. We know that the backwards rationalizability procedure delivers exactly the strategies that can rationally be chosen under common belief in future rationality. In the previous subsection we have seen that the order of elimination does not matter, so we may as well use the backward induction order described above. So, we first consider all information sets at the end of the game, at which the backwards rationalizability procedure deletes all suboptimal choices. That is, we uniquely select the backward induction choices at all information sets at the end of the game.

Next, we turn to the information sets just before these, where we start by deleting the strategies that were already deleted at the previous round. In this case, we would thus keep only those strategies that prescribe the backward induction choice at the last information sets in the game. Then, we would delete those strategies that are not optimal against the surviving strategies, that is, we remove strategies that are not optimal against the backward induction choices at the end of the game. So, we select the backward induction choices also at information sets just before the last information sets in the game.

By iterating this argument, we see that applying the backwards rationalizability procedure in the backward induction fashion would yield exactly the backward induction choice at every information set. Consequently, we obtain the unique backward induction strategy for every player. Since the order of elimination does not matter, as we have seen, we conclude that applying the backwards rationalizability procedure to a game with perfect information and no relevant ties would yield precisely the backward induction strategies for the players.

Together with Theorem 4.3 we thus see that in every game with perfect information and no relevant ties, common belief in future rationality leads to backward induction.

Theorem 6.1 *(Common belief in future rationality leads to backward induction). Consider a dynamic game with perfect information and no relevant ties. Then, every player has exactly one strategy he can rationally choose under common belief in future rationality, namely his backward induction strategy.*

So we see that the order independence of the backwards rationalizability procedure can also be used to provide relationships between common belief in future rationality and other concepts in the literature.

6.5. *Best-response characterization*

We will finally use the backwards rationalizability procedure to provide a characterization of common belief in future rationality in terms of "best responses". For every information set h, let $S_i^\infty(h)$ be the set of strategies for player i that survive the backwards rationalizability procedure at h. Moreover, let B_i^∞ be the set of conditional belief vectors selected by the procedure at the end. By construction of the procedure, these sets $S_i^\infty(h)$ and B_i^∞ have the following properties:

(1) $S_i^\infty(h)$ contains precisely those strategies $s_i \in S_i(h)$ for player i that are optimal, for some $b_i \in B_i^\infty$, at every $h_i' \in H_i(s_i)$ weakly following h.

(2) B_i^∞ contains precisely those conditional belief vectors b_i that satisfy Bayesian updating, and for which $b_i(h_i) \in \Delta(S_{-i}^\infty(h_i))$ at every $h_i \in H_i$.

By combining (1) and (2), we obtain the following characterization of the sets $S_i^\infty(h)$:

$S_i^\infty(h)$ contains precisely those strategies $s_i \in S_i(h)$ for player i that are optimal, at every $h_i' \in H_i(s_i)$ weakly following h, for some conditional belief vector b_i that satisfies Bayesian updating, and for which $b_i(h_i') \in \Delta(S_{-i}^\infty(h_i'))$ at every $h_i' \in H_i$ weakly following h.

We say that the collection $(S_i^\infty(h))_{h\in H, i\in I}$ of strategy sets is "closed under belief in future rationality". Here, H denotes the collection of all information sets. We first formally define what we mean by "closed under belief in future rationality".

Definition 6.2 *(Closed under belief in future rationality).* For every information set h, and every player i, let $D_i(h) \subseteq S_i(h)$ be some subset of strategies. The collection $(D_i(h))_{h\in H, i\in I}$ of strategy subsets is closed under belief in future rationality if for every $s_i \in D_i(h)$ there is some belief vector b_i satisfying Bayesian updating, such that (1) $b_i(h_i') \in \Delta(D_{-i}(h_i'))$ for every $h_i' \in H_i$ weakly following h, and (2) s_i is optimal for $b_i(h_i')$ at every $h_i' \in H_i(s_i)$ weakly following h.

We now show that the strategies that can rationally be chosen under common belief in future rationality are exactly those that correspond to some collection of strategy subsets which is closed under belief in future rationality.

Theorem 6.3 *(Best-response characterization of common belief in future rationality). A strategy s_i can rationally be chosen under common belief in future rationality, if and only if, there is a collection $(D_i(h))_{h\in H, i\in I}$ of strategy subsets which is closed under belief in future rationality, and in which $s_i \in D_i(\emptyset)$.*

The proof can be found in Section 9. In fact, the proof tells us a little bit more, namely that the collection $(S_i^\infty(h))_{h\in H, i\in I}$ of strategy subsets surviving the backwards rationalizability procedure is the *largest* collection that is closed under belief in future rationality. In general, there may be other, smaller collections which are also closed under belief in future rationality.

7. Relation to other concepts

In this section we will investigate the relation that common belief in future rationality bears with other epistemic concepts for dynamic games, in particular sequential rationalizability and extensive form rationalizability.

7.1. Sequential rationalizability

The concept of *sequential rationalizability* has been proposed independently by Dekel et al. (1999, 2002) (DFL from now on) and Asheim and Perea (2005), although they differ considerably in their formulation. Here we will use the formulation by DFL as it makes it easier to compare the concept to our notion of common belief in future rationality. The key ingredients in DFL's model are

(a) *behavioral strategies* π_i, which assign to every information set h_i for player i a probability distribution over i's choices at h. A behavioral strategy π_i represents i's strategy choice;
(b) *assessments* a_i, which assign to every information set h_i for player i a probability distribution over the histories in h. An assessment a_i represents i's conditional beliefs about the opponents' *past* behavior; and
(c) profiles π_{-i}^i of *behavioral strategies* for i's opponents. A profile π_{-i}^i represents i's conditional beliefs about the opponents' *future* behavior.

Since the profile π_{-i}^i is not correlated, the last ingredient implies that player i's belief about opponent j's future behavior should be independent from his belief about opponent k's future behavior. A conditional belief pair (a_i, π_{-i}^i) is called *Kreps–Wilson consistent* (Kreps and Wilson, 1982) if there is a sequence $(a_i^n, \pi_{-i}^{i,n})_{n\in\mathbb{N}}$ converging to (a_i, π_{-i}^i) in which $\pi_{-i}^{i,n}$ assigns positive probability to all choices, and a_i^n is obtained from $\pi_{-i}^{i,n}$ by Bayesian updating.

For every player i, consider a set V_i of strategy-belief triples (π_i, a_i, π_{-i}^i). The collection $V = (V_i)_{i\in I}$ of sets of strategy-belief triples is called *sequentially rationalizable* if for every $(\pi_i, a_i, \pi_{-i}^i) \in V_i$,

(a) (a_i, π_{-i}^i) is Kreps–Wilson consistent,
(b) strategy π_i is optimal at every information set $h_i \in H_i$ under the belief (a_i, π_{-i}^i), and
(c) the belief π_{-i}^i about the opponents' future behavior only assigns positive probability to opponents' strategies π_j which are part of some triple in V.[3]

The last two conditions together thus state that a player, at every information set, should only assign positive probability to opponents' strategies that, at every *future* information set, are optimal for some belief in V. Finally, a strategy π_i is called

[3] For a precise statement of this condition, see Definition 2.2 in Dekel et al. (2002).

sequentially rationalizable if there is some sequentially rationalizable collection $(V_i)_{i\in I}$ of sets of strategy-belief triples, such that π_i is part of some triple in V_i.

Let us now try to translate this concept in terms of conditional beliefs as we use them in this paper. The conditional belief pair (a_i, π^i_{-i}) in DFL corresponds to a conditional belief vector $(b_i(h_i))_{h_i\in H_i}$ in our setup, where $b_i(h_i)$ is a probability distribution over $S_{-i}(h_i)$ for every $h_i \in H_i$. This conditional belief vector $(b_i(h_i))_{h_i\in H_i}$ should be such, however, that i's conditional belief at h_i about the opponents' future behavior is independent across opponents. For every player i, consider a set $\tilde{V}_i$ of conditional belief vectors $(b_i(h_i))_{h_i\in H_i}$. Then, the collection $\tilde{V} = (\tilde{V}_i)_{i\in I}$ is *sequentially rationalizable* if for every $(b_i(h_i))_{h_i\in H_i} \in \tilde{V}_i$,

(d) at every $h_i \in H_i$, the conditional belief about the opponents' future behavior is independent across opponents,
(e) the conditional belief vector $(b_i(h_i))_{h_i\in H_i}$ is Kreps–Wilson consistent,
(f) at every $h_i \in H_i$, the conditional belief $b_i(h_i)$ only assigns positive probability to opponents' strategies s_j which are optimal, at every $h'_j \in H_j(s_j)$ *weakly following* h_i, for some conditional belief vector in $\tilde{V}_j$.

Here, condition (f) follows from our insight above that in DFL's definition, a player should, at every information set, only assign positive probability to opponents' strategies that, at every *future* information set, are optimal for some belief in V_j. So, a strategy s_i is sequentially rationalizable, if and only if, there is some sequentially rationalizable collection $(\tilde{V}_i)_{i\in I}$ of conditional belief vectors, and some conditional belief vector in $\tilde{V}_i$, for which s_i is optimal at every information set.

Now, take a sequentially rationalizable collection $(\tilde{V}_i)_{i\in I}$ of conditional belief vectors. For every player i, every information set $h_i \in H_i$, and every opponent j, let $D_j(h_i) \subseteq S_j(h_i)$ be the set of strategies that receive positive probability at h_i under some conditional belief in $\tilde{V}_i$. At an information set $h_i \in H_i$, let $D_i(h_i) \subseteq S_i(h_i)$ be the set of strategies in $S_i(h_i)$ that are optimal, at every $h'_i \in H_i$ weakly following h_i, for some conditional belief vector in $\tilde{V}_i$.

Note that every conditional belief vector that is Kreps–Wilson consistent, also satisfies Bayesian updating. But then, by conditions (e) and (f) above, we know that the collection $(D_i(h))_{i\in I, h\in H}$ of strategy subsets has the following property:

If $s_i \in D_i(h)$, then s_i is optimal, at every $h'_i \in H_i(s_i)$ weakly following h, for some conditional belief vector b_i that (1) satisfies Bayesian updating and (2) for which $b_i(h'_i) \in \Delta(D_{-i}(h'_i))$ at every $h'_i \in H_i$ weakly following h. That is, the collection $(D_i(h))_{i\in I, h\in H}$ is closed under belief in future rationality, conform our Definition 6.2. We have thus shown that every sequentially rationalizable collection $(\tilde{V}_i)_{i\in I}$ of conditional belief vectors induces, in a natural way, a collection $(D_i(h))_{i\in I, h\in H}$ of strategy subsets that is closed under belief in future rationality. But then, it immediately follows from our Theorem 6.3 that every sequentially rationalizable strategy can rationally be chosen under common belief in future rationality. We have thus established the following result.

Theorem 7.1 *(Relation to sequential rationalizability). Every sequentially rationalizable strategy can rationally be chosen under common belief in future rationality.*

The opposite direction is not true in general. Consider, namely, a one-shot game with three players or more, in which all players simultaneously make one choice. For such games, the concept of sequential rationalizability coincides with *uncorrelated rationalizability*, as defined by Bernheim (1984) and Pearce (1984), in which it is assumed that a player's belief about the opponents' choice combinations is independent across opponents. Moreover, for these games our concept of common belief in future rationality coincides with the concept of *correlated rationalizability*, in which a player's belief about the opponents' choice combinations may be correlated across opponents. It is well-known that for such games, uncorrelated rationalizability may be strictly more restrictive than correlated rationalizability. Hence, in general, the concept of sequential rationalizability may be strictly more restrictive than common belief in future rationality.

The theorem above also implies that common belief in future rationality is always possible in every dynamic game, as sequentially rational strategies always exist (see for instance Asheim and Perea, 2005). We thus obtain the following existence result.

Theorem 7.2 *(Common belief in future rationality is always possible). For every finite dynamic game Γ there is an epistemic model $M = (T_i, b_i)_{i\in I}$ such that for every player i there is some type $t_i \in T_i$ that expresses common belief in future rationality.*

In Asheim and Perea (2005) it is shown that sequential equilibrium constitutes a refinement of sequential rationalizability. More precisely, every strategy that receives positive probability in a sequential equilibrium is sequentially rationalizable. From Theorem 7.1 it thus follows that every strategy that receives positive probability in a sequential equilibrium can rationally be chosen under common belief in future rationality. That is, common belief in future rationality is a weakening of the concept of sequential equilibrium.

7.2. Extensive form rationalizability

The concept of *extensive form rationalizability* has originally been proposed in Pearce (1984) by means of an iterated reduction procedure. Later, Battigalli (1997) has simplified this procedure and has shown that it delivers the same output as

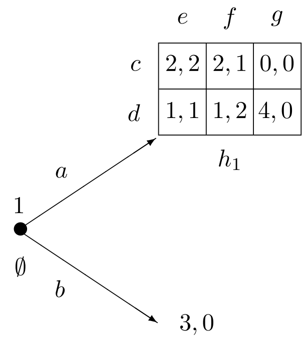


Fig. 4. There is no logical relationship, in terms of strategies, between common belief in future rationality and extensive form rationalizability.

Pearce's procedure. More recently, Battigalli and Siniscalchi (2002) have provided an epistemic characterization of extensive form rationalizability – *common strong belief in rationality*.

In this section we wish to compare our notion of common belief in future rationality to the concept of extensive form rationalizability. A more detailed comparison of these two concepts can be found in Perea (2010), which analyzes the differences between backward and forward induction reasoning in dynamic games. To carry out this comparison we will use yet another procedure leading to extensive form rationalizability, namely the *iterated conditional dominance* procedure developed by Shimoji and Watson (1998). The reason is that this procedure is closer to our backward dominance algorithm, and therefore easier to compare.

Indeed, Shimoji and Watson's procedure is very similar in spirit to our backward dominance procedure, as it iteratively removes strategies from decision problems. However, their criterion for removing a strategy in a particular decision problem is different. Formally, their procedure can be formulated as follows.

Algorithm 7.3 *(Shimoji and Watson's iterated conditional dominance procedure).*

Initial step. *For every information set h, define* $\Gamma^0(h) := S(h)$.

Inductive step. *Let* $k \geqslant 1$, *and suppose that the decision problems* $\Gamma^{k-1}(h)$ *have been defined for every information set h. Then, at every information set h we define*

$$\Gamma^k(h) := \Gamma^{k-1}(h) \backslash \bigcup_{h' \in H} sd\big(\Gamma^{k-1}(h')\big).$$

A strategy s_i *survives the iterated conditional dominance procedure if there is some* $s_{-i} \in S_{-i}$ *such that* $(s_i, s_{-i}) \in \Gamma^k(\emptyset)$ *for all k.*

Remember from the backward dominance procedure that $sd(\Gamma^{k-1}(h'))$ contains all those strategy combinations $(s_i)_{i\in I}$ in $\Gamma^{k-1}(h')$ such that s_i is strictly dominated within $\Gamma^{k-1}(h')$ for some active player i at h'. By H we denote the collection of all information sets in the game.

So, the crucial difference with the backward dominance procedure is that in the iterated conditional dominance procedure, we eliminate a strategy from $\Gamma^{k-1}(h)$ whenever it is strictly dominated at some decision problem $\Gamma^{k-1}(h')$ that either precedes h, or weakly follows h, or neither precedes nor follows h. In contrast, within the backward dominance procedure we only eliminate a strategy from $\Gamma^{k-1}(h)$ if it is strictly dominated at some decision problem that *weakly follows* h.

Shimoji and Watson (1998) have shown that the iterated conditional dominance procedure delivers exactly the set of extensive form rationalizable strategies. Moreover, they prove that for the output of the iterated conditional dominance procedure it is inessential whether we impose (common belief in) Bayesian updating or not. So, the iterated conditional dominance procedure in combination with common belief in Bayesian updating would also yield exactly the extensive form rationalizable strategies.

Note that in the iterated conditional dominance procedure, it is possible that at a given decision problem $\Gamma^{k-1}(h)$ *all* strategy combinations will be eliminated in step k – something that can never happen in the backward dominance procedure. In other words, the decision problem $\Gamma^k(h)$ may become empty at some step k. Consider, namely, some information set $h_i \in H_i$, and some information set h' following h_i. Then, it is possible that within the decision problem $\Gamma^{k-1}(h_i)$, all strategies for player i that are in $\Gamma^{k-1}(h')$ are strictly dominated within $\Gamma^{k-1}(h_i)$. In that case, we would eliminate in $\Gamma^{k-1}(h')$ all remaining strategies for player i, and hence all remaining strategy combinations!

To illustrate this important aspect, let us consider the game from Fig. 2, and replace the utilities 4, 0 after choice b by 7, 0. Then, in the first step of the iterated conditional dominance procedure we would eliminate strategies (a, c), (a, d) and (a, e) for player 1 at h_1, as they are all strictly dominated by b at $\emptyset$. So, after Step 1 we have no strategy combinations left at h_1 – that is, $\Gamma^1(h_1)$ would be empty.

Now, what can we say about the relationship between common belief in future rationality and extensive form rationalizability? It turns out that in terms of *strategies*, there is no logical relationship between the two concepts. Consider, to that purpose, the game in Fig. 4. The full decision problems at $\emptyset$ and h_1 are represented in Table 6.

Table 6
The full decision problems in Fig. 3:

Player 1 active

$\Gamma^0(\emptyset)$	e	f	g
(a,c)	2, 2	2, 1	0, 0
(a,d)	1, 1	1, 2	4, 0
b	3, 0	3, 0	3, 0

Players 1 and 2 active

$\Gamma^0(h_1)$	e	f	g
(a,c)	2, 2	2, 1	0, 0
(a,d)	1, 1	1, 2	4, 0

The backward dominance procedure does the following: In the first round, we eliminate from $\Gamma^0(\emptyset)$ strategy (a,c) as it is strictly dominated by b at player 1's decision problem $\Gamma^0(\emptyset)$, and we eliminate from $\Gamma^0(\emptyset)$ and $\Gamma^0(h_1)$ strategy g as it is strictly dominated by e and f at player 2's decision problem $\Gamma^0(h_1)$. In the second round, we eliminate from $\Gamma^1(\emptyset)$ strategy (a,d) as it strictly dominated by b at $\Gamma^1(\emptyset)$, and we eliminate strategy (a,d) also from $\Gamma^1(h_1)$ as it is strictly dominated by (a,c) at $\Gamma^1(h_1)$. In the third round, finally, we eliminate from $\Gamma^2(\emptyset)$ and $\Gamma^2(h_1)$ strategy f, as it is strictly dominated by e in $\Gamma^2(h_1)$. So, only strategies b and e remain at $\emptyset$. Hence, the backward dominance procedure uniquely yields the strategies b and e for players 1 and 2, respectively. But then, by Theorem 5.4, only strategies b and e can rationally be chosen under common belief in future rationality.

The iterated conditional dominance procedure works differently here: In the first round, we eliminate strategy (a,c) from $\Gamma^0(\emptyset)$ *and* $\Gamma^0(h_1)$ as it is strictly dominated by b at player 1's decision problem $\Gamma^0(\emptyset)$, and we eliminate from $\Gamma^0(\emptyset)$ and $\Gamma^0(h_1)$ strategy g as it is strictly dominated by e and f at player 2's decision problem $\Gamma^0(h_1)$. In the second round, we eliminate (a,d) from $\Gamma^1(\emptyset)$ and $\Gamma^1(h_1)$ as it is strictly dominated by b at $\Gamma^1(\emptyset)$ – thus making $\Gamma^2(h_1)$ empty – and we eliminate e from $\Gamma^1(\emptyset)$ and $\Gamma^1(h_1)$ as it is strictly dominated by f in $\Gamma^1(h_1)$. This only leaves strategies b and f at $\emptyset$, and hence only strategies b and f can be chosen under extensive form rationalizability.

In particular, we see that common belief in future rationality uniquely selects strategy e for player 2, whereas extensive form rationalizability singles out strategy f for player 2. The crucial difference lies in how player 2 at h_1 explains the surprise that player 1 has not chosen b. Under common belief in future rationality, player 2 believes at h_1 that player 1 has simply made a mistake, but he still believes that player 1 will choose rationally at h_1, and he still believes that player 1 believes that player 2 will not choose g at h_1. So, player 2 believes at h_1 that player 1 will choose (a,c), and therefore player 2 will choose e at h_1. Under extensive form rationalizability, player 2 believes at h_1 that player 1's decision not to choose b was a rational decision, but this is only possible if player 2 believes at h_1 that player 1 believes that player 2 will irrationally choose g at h_1. In that case, player 2 will believe at h_1 that player 1 will go for (a,d), and therefore player 2 will choose f at h_1.

The game in Fig. 4 thus shows that, in terms of strategies, common belief in future rationality and extensive form rationalizability may yield unique but opposite predictions for a certain player. Note, however, that in this game both concepts lead to the same *outcome*, namely b.

This leads to the following question: Is it possible to find games where both concepts would also yield unique but different *outcomes*? The answer is "no". In Chapter 9 of Perea (2012) it is shown, namely, that every outcome which can be realized under extensive form rationalizability, can also be realized under common belief in future rationality.

This result also follows from Chen and Micali (2013). They show, namely, that changing the order of elimination in the iterated conditional dominance procedure does not change the *outcomes* that are selected by the procedure – although it may change the *strategies* selected. Now, it can be verified that the backward dominance procedure corresponds to the *first few steps* in the iterated conditional dominance procedure – by choosing a very specific, different order of elimination – but without necessarily completing the procedure after these first few steps! By combining these two facts, we thus conclude that the outcomes selected by the backward dominance procedure will always *contain* the outcomes selected by the iterated conditional dominance procedure. As common belief in future rationality corresponds to the backward dominance procedure in combination with common belief in Bayesian updating, and extensive form rationalizability corresponds to the iterated conditional dominance procedure in combination with common belief in Bayesian updating, it follows that, in terms of outcomes, the concept of extensive form rationalizability is more restrictive than common belief in future rationality.

8. Future research

A possibly interesting application of the idea of common belief in future rationality would be to investigate its behavioral implications for finitely and infinitely repeated games. Although infinitely repeated games fall outside the class of games considered in this paper, the concept of common belief in future rationality could be defined for such games as well. A question that could be addressed is: Can we find an easy algorithm that computes, for every stage of the repeated game, the set of choices a player can make there under common belief in future rationality? As a next step, one could also explore the idea of common belief in future rationality in discounted stochastic games, which include finitely and infinitely repeated games as special cases. An interesting question, similar to the one above, would be: Is there an algorithm that computes, for every state, the set of choices a player can make there under common belief in future rationality?

9. Proofs

In this section we will deliver formal proofs for the theorems in this paper. Before doing so, we first present some preparatory results that will play a crucial role in some of these proofs.

9.1. Some preparatory results

For a given player i, let $(D_{-i}(h_i))_{h_i \in H_i}$ be a collection of nonempty strategy subsets $D_{-i}(h_i) \subseteq S_{-i}(h_i)$. Say that $(b_i(h_i))_{h_i \in H_i}$ is a *conditional belief vector on* $(D_{-i}(h_i))_{h_i \in H_i}$ if $b_i(h_i) \in \Delta(D_{-i}(h_i))$ for every $h_i \in H_i$. Fix some information set $h_i^* \in H_i$, and some conditional belief $b_i(h_i^*) \in \Delta(D_{-i}(h_i^*))$. The question is: Can we extend $b_i(h_i^*)$ to a conditional belief vector $(b_i(h_i))_{h_i \in H_i}$ on $(D_{-i}(h_i))_{h_i \in H_i}$ such that there exists a strategy $s_i \in S_i(h_i^*)$ which is optimal, at every $h_i \in H_i$ weakly following h_i^*, for the belief $b_i(h_i)$? We provide a sufficient condition under which this is indeed possible.

Definition 9.1 *(Forward inclusion property).* The collection $(D_{-i}(h_i))_{h_i \in H_i}$ of strategy subsets $D_{-i}(h_i) \subseteq S_{-i}(h_i)$ satisfies the forward inclusion property if for every $h_i, h_i' \in H_i$ where h_i' follows h_i, it holds that $D_{-i}(h_i) \cap S_{-i}(h_i') \subseteq D_{-i}(h_i')$.

Lemma 9.2 *(Existence of sequentially optimal strategies). For a given player i, consider a collection $(D_{-i}(h_i))_{h_i \in H_i}$ of strategy subsets satisfying the forward inclusion property. At a given information set $h_i^* \in H_i$ fix some conditional belief $b_i(h_i^*) \in \Delta(D_{-i}(h_i^*))$. Then, $b_i(h_i^*)$ can be extended to a conditional belief vector $(b_i(h_i))_{h \in H_i}$ on $(D_{-i}(h_i))_{h_i \in H_i}$, such that there is some strategy $s_i \in S_i(h_i^*)$ which is optimal at every $h_i \in H_i$ weakly following h_i^* for the belief $b_i(h_i)$.*

Proof. Fix some information set $h_i^* \in H_i$, and some conditional belief $b_i(h_i^*) \in \Delta(D_{-i}(h_i^*))$. We will extend $b_i(h_i^*)$ to some conditional belief vector $(b_i(h_i))_{h_i \in H_i}$ on $(D_{-i}(h_i))_{h_i \in H_i}$, and construct some strategy $s_i \in S_i(h_i^*)$, such that s_i is optimal at every $h_i \in H_i$ weakly following h_i^* for the belief $b_i(h_i)$.

Let $H_i(h_i^*)$ be the collection of player i information sets that follow h_i^*. Let $H_i^+(h_i^*)$ be those information sets $h_i \in H_i(h_i^*)$ with $b_i(h_i^*)(S_{-i}(h_i)) > 0$, where $b_i(h_i^*)(S_{-i}(h_i))$ is a short way to write $\sum_{s_{-i} \in S_{-i}(h_i)} b_i(h_i^*)(s_{-i})$. For every $h_i \in H_i^+(h_i^*)$ we define the conditional belief $b_i(h_i) \in \Delta(D_{-i}(h_i))$ by

$$b_i(h_i)(s_{-i}) := \frac{b_i(h_i^*)(s_{-i})}{b_i(h_i^*)(S_{-i}(h_i))}$$

for every $s_{-i} \in S_{-i}(h_i)$. So, $b_i(h_i)$ is obtained from $b_i(h_i^*)$ by Bayesian updating. To see that $b_i(h_i) \in \Delta(D_{-i}(h))$, note that $b_i(h_i)$ only assigns positive probability to $s_{-i} \in S_{-i}(h_i)$ that received positive probability under $b_i(h_i^*)$. Since, by construction, $b_i(h_i^*) \in \Delta(D_{-i}(h_i^*))$, it follows that $b_i(h_i)$ only assigns positive probability to $s_{-i} \in D_{-i}(h_i^*) \cap S_{-i}(h_i)$. However, by the forward inclusion property, $D_{-i}(h_i^*) \cap S_{-i}(h_i) \subseteq D_{-i}(h_i)$, and hence $b_i(h_i) \in \Delta(D_{-i}(h_i))$.

Now, consider an information set $h_i \in H_i(h_i^*) \backslash H_i^+(h_i^*)$ which is not preceded by any $h_i' \in H_i(h_i^*) \backslash H_i^+(h_i^*)$. That is, $b_i(h_i^*)(S_{-i}(h_i)) = 0$, but $b_i(h_i^*)(S_{-i}(h_i')) > 0$ for every $h_i' \in H_i$ between h_i^* and h_i. For every such h_i, choose some arbitrary conditional belief $b_i(h_i) \in \Delta(D_{-i}(h_i))$.

Let $H_i^+(h_i)$ be those information sets $h_i' \in H_i$ weakly following h_i with $b_i(h_i)(S_{-i}(h_i')) > 0$. For every $h_i' \in H_i^+(h_i)$, define the conditional belief $b_i(h_i')$ as above, so $b_i(h_i')$ is obtained from $b_i(h_i)$ by Bayesian updating. By the same argument as above, it can be shown that $b_i(h_i') \in \Delta(D_{-i}(h_i'))$ for every $h_i' \in H_i^+(h_i)$.

By continuing in this fashion, we will finally define for every $h_i \in H_i$ following h_i^* some conditional belief $b_i(h_i) \in \Delta(D_{-i}(h_i))$, such that these conditional beliefs, together with $b_i(h_i^*)$, satisfy Bayesian updating where possible. For every information set $h_i \in H_i$ not weakly following h_i^*, define $b_i(h_i) \in \Delta(D_{-i}(h_i))$ arbitrarily. So, $(b_i(h_i))_{h_i \in H_i}$ is a conditional belief vector on $(D_{-i}(h_i))_{h_i \in H_i}$ which extends $b_i(h_i^*)$, and it satisfies Bayesian updating at information sets weakly following h_i^*.

We will now construct a strategy $s_i \in S_i(h_i^*)$ that, at every $h_i \in H_i$ weakly following h_i^*, is optimal for the belief $b_i(h_i)$. By "backward induction", we choose at every $h_i \in H_i$ weakly following h_i^* a choice $c_i(h_i) \in C_i(h_i)$ that is optimal at h_i for the belief $b_i(h_i)$, given player i's own choices at future histories. More precisely, we start with information sets $h_i \in H_i$ weakly following h_i^* which are not followed by any other player i information set. At those h_i, we specify a choice $c_i(h_i) \in C_i(h_i)$ with

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

Now, suppose that $h_i \in H_i$ weakly follows h_i^*, and that $c_i(h_i')$ has been defined for all $h_i' \in H_i$ following h_i. Then, we specify a choice $c_i(h_i) \in C_i(h_i)$ with

$$u_i\Big(\big(c_i(h_i), \big(c_i\big(h_i'\big)\big)_{h_i' \in H_i(h)}\big), b_i(h_i)\Big) \geqslant u_i\Big(\big(c_i', \big(c_i\big(h_i'\big)\big)_{h_i' \in H_i(h)}\big), b_i(h_i)\Big) \tag{9.1}$$

for all $c_i' \in C_i(h_i)$. Here, $H_i(h_i)$ denotes the collection of information sets in H_i that follow h_i. In this way, we specify at every $h_i \in H_i$ weakly following h_i^* a choice $c_i(h_i)$ that satisfies (9.1).

Now, let s_i be the strategy that

(a) at every $h_i \in H_i(s_i)$ weakly following h_i^*, prescribes the optimal choice $c_i(h_i)$ as in (9.1),
(b) at every $h_i \in H_i(s_i)$ preceding h_i^*, prescribes the unique choice $c_i(h_i)$ that leads to h_i^*, and
(c) at every other $h_i \in H_i(s_i)$ specifies an arbitrary choice.

By construction the strategy s_i is in $S_i(h_i^*)$, as it prescribes all choices that lead to h_i^*. As the conditional belief vector $(b_i(h_i))_{h_i \in H_i}$ satisfies Bayesian updating at information sets weakly following h_i^*, it follows from Theorem 3.1 in Perea (2002) that this profile of beliefs satisfies the *one-deviation property* at information sets weakly following h_i^*. That is, every strategy s_i for which the choices $c_i(h_i)$ are optimal in the sense of (9.1), is optimal as a strategy at every $h_i \in H_i$ weakly following h_i^*. Hence, we may conclude that the strategy s_i so constructed is optimal at every $h_i \in H_i(s_i)$ weakly following h_i^* for the belief $b_i(h_i)$. Since s_i is in $S_i(h_i^*)$, and $(b_i(h_i))_{h_i \in H_i}$ is a conditional belief vector on $(D_{-i}(h_i))_{h_i \in H_i}$ which extends $b_i(h_i^*)$, the proof is complete. □

The lemma above implies in particular that, whenever the collection $(D_{-i}(h_i))_{h_i \in H_i}$ satisfies the forward inclusion property, then it allows for a conditional belief vector $(b_i(h_i))_{h_i \in H_i}$ and a strategy s_i, such that s_i is optimal at every $h_i \in H_i(s_i)$ for the belief $b_i(h_i)$. In other words, collections $(D_{-i}(h_i))_{h_i \in H_i}$ that satisfy the forward inclusion property allow for strategies that are sequentially optimal. We believe this an interesting result which may be useful for other applications as well.

Our second result shows that the sets of strategies surviving a particular round of the backward dominance procedure satisfy the forward inclusion property. This result thus guarantees that we can apply Lemma 9.2 to every round of the backward dominance procedure – something that will be important for proving some of our theorems in the paper.

Lemma 9.3 *(Backward dominance procedure satisfies forward inclusion property). For every player i and information set $h_i \in H_i$, let $\Gamma^k(h_i) = S_i^k(h_i) \times S_{-i}^k(h_i)$, where $\Gamma^k(h_i)$ is the decision problem at h_i produced in round k of the backward dominance procedure. Then, the collection $(S_{-i}^k(h_i))_{h \in H_i}$ of strategy subsets satisfies the forward inclusion property.*

Proof. For $k = 0$ the statement is trivial since $S_{-i}^0(h_i) = S_{-i}(h_i)$ for all $h_i \in H_i$. So, take some $k \geqslant 1$. Suppose that $h_i, h_i' \in H_i$ and that h_i' follows h_i. Take some opponent's strategy combination s_{-i} in $S_{-i}^k(h_i) \cap S_{-i}(h_i')$, where $s_{-i} = (s_j)_{j \neq i}$. Then, for every $j \neq i$, we have that s_j is not strictly dominated in any decision problem $\Gamma^{k-1}(h_j'')$ where $h_j'' \in H_j(s_j)$ weakly follows h_i. As h_i' follows h_i, it holds in particular that s_j is not strictly dominated in any decision problem $\Gamma^{k-1}(h_j'')$ where $h_j'' \in H_j(s_j)$ weakly follows h_i'. Together with the fact that $s_{-i} \in S_{-i}(h_i')$, this implies that $s_{-i} \in S_{-i}^k(h_i')$. So, $S_{-i}^k(h_i) \cap S_{-i}(h_i') \subseteq S_{-i}^k(h_i')$, and hence the forward inclusion property holds. □

Our third lemma shows an important optimality property of our backward dominance procedure. Recall that in the backward dominance procedure, $\Gamma^k(h)$ denotes the decision problem at h produced at the end of round k. We say that $\Gamma^k(h)$ contains a strategy s_i for player i if there is some $s_{-i} \in S_{-i}$ such that $(s_i, s_{-i}) \in \Gamma^k(h)$. For every player i, let $S_i^k(h)$ be the set of strategies that are contained in $\Gamma^k(h)$. In particular, if player i is active at h then we have that $\Gamma^k(h) = S_i^k(h) \times S_{-i}^k(h)$ for some $S_{-i}^k(h) \subseteq S_{-i}(h)$.

By construction of the algorithm, $S_i^k(h)$ contains exactly those strategies in $S_i^{k-1}(h)$ that, at every $h_i' \in H_i(s_i)$ weakly following h, are not strictly dominated in $\Gamma^{k-1}(h_i')$. By Lemma 3 in Pearce (1984), we know that s_i is not strictly dominated in $\Gamma^{k-1}(h_i')$ if and only if there is some belief $b_i(h_i') \in \Delta(S_{-i}^{k-1}(h_i'))$ such that s_i is optimal for $b_i(h_i')$ *among all strategies in* $S_i^{k-1}(h_i')$. That is,

$$u_i\bigl(s_i, b_i\bigl(h_i'\bigr)\bigr) \geqslant u_i\bigl(s_i', b_i\bigl(h_i'\bigr)\bigr) \quad \text{for all } s_i' \in S_i^{k-1}\bigl(h_i'\bigr).$$

However, we can show a little more about s_i: Not only is s_i optimal for the belief $b_i(h_i')$ among all strategies in $S_i^{k-1}(h_i')$, it is even optimal among *all strategies in* $S_i(h_i')$. That is, at every $h_i' \in H_i(s_i)$ weakly following h we even have that

$$u_i\bigl(s_i, b_i\bigl(h_i'\bigr)\bigr) \geqslant u_i\bigl(s_i', b_i\bigl(h_i'\bigr)\bigr) \quad \text{for all } s_i' \in S_i\bigl(h_i'\bigr).$$

We call this the *optimality principle* for the backward dominance procedure, and it will play a crucial role in proving some of the results in our paper.

Lemma 9.4 *(Optimality principle for backward dominance procedure). Let $S_i^k(h)$ denote the set of player i strategies contained in the decision problem $\Gamma^k(h)$ produced in round k of the backward dominance procedure. Then, $s_i \in S_i^k(h)$ if and only if for every $h_i' \in H_i(s_i)$ weakly following h there is some belief $b_i(h_i') \in \Delta(S_{-i}^{k-1}(h_i'))$ such that s_i is optimal for $b_i(h_i')$ among all strategies in $S_i(h_i')$.*

Proof. The "if" direction follows immediately, so we only have to prove the "only if" direction. Fix some information set h, some player i, some strategy $s_i \in S_i^k(h)$, and some $h_i' \in H_i(s_i)$ weakly following h. Then we know from our argument above that there is some $b_i(h_i') \in \Delta(S_{-i}^{k-1}(h_i'))$ such that

$$u_i(s_i, b_i(h_i')) \geqslant u_i(s_i', b_i(h_i')) \quad \text{for all } s_i' \in S_i^{k-1}(h_i'). \tag{9.2}$$

We will prove that, in fact,

$$u_i(s_i, b_i(h_i')) \geqslant u_i(s_i', b_i(h_i')) \quad \text{for all } s_i' \in S_i(h_i').$$

Suppose, on the contrary, that there would be some $s_i' \in S_i(h_i')$ such that

$$u_i(s_i, b_i(h_i')) < u_i(s_i', b_i(h_i')). \tag{9.3}$$

We show that in this case there would be some $s_i^* \in S_i^{k-1}(h_i')$ with $u_i(s_i', b_i(h_i')) \leqslant u_i(s_i^*, b_i(h_i'))$, which together with (9.3) would contradict (9.2).

From Lemma 9.3 we know that the collection $(S_{-i}^{k-1}(h_i''))_{h_i'' \in H_i}$ satisfies the forward inclusion property. Hence, by Lemma 9.2, we can extend $b_i(h_i')$ to some conditional belief vector $(b_i(h_i''))_{h_i'' \in H_i}$ with $b_i(h_i'') \in \Delta(S_{-i}^{k-1}(h_i''))$ for all $h_i'' \in H_i$, and we can find some strategy $s_i^* \in S_i(h_i')$ which is optimal, at every $h_i'' \in H_i(s_i^*)$ weakly following h_i', for the belief $b_i(h_i'')$. But then, it follows that $s_i^* \in S_i^k(h_i')$, and hence in particular $s_i^* \in S_i^{k-1}(h_i')$. Moreover, s_i^* is optimal at h_i' for the belief $b_i(h_i')$. And hence, we have by (9.3) that

$$u_i(s_i, b_i(h_i')) < u_i(s_i', b_i(h_i')) \leqslant u_i(s_i^*, b_i(h_i')) \quad \text{for some } s_i^* \in S_i^{k-1}(h_i').$$

This, however, contradicts (9.2). So, (9.3) must be incorrect, and hence s_i is optimal at h_i' for the belief $b_i(h_i')$ among all strategies in $S_i(h_i')$. □

9.2. Every $\Gamma^k(h)$ in the backward dominance procedure is a decision problem

We now prove Lemma 5.2, which states that every $\Gamma^k(h)$ within the backward dominance procedure is a decision problem. That is, we must show that for every $k \geqslant 0$, every information set h, and every active player i at h, there are sets $D_i \subseteq S_i(h)$ and $D_{-i} \subseteq S_{-i}(h)$ such that $\Gamma^k(h) = D_i \times D_{-i}$. We prove this statement by induction on k.

Consider first the case $k = 0$. Take some information set h and some active player i at h. By definition, $\Gamma^0(h) = S(h)$. Moreover, by perfect recall we have that $S(h) = S_i(h) \times S_{-i}(h)$, and hence $\Gamma^0(h) = S_i(h) \times S_{-i}(h)$. So the statement holds for $k = 0$.

Take now some $k \geqslant 1$, and assume that the statement holds for $k - 1$. Take some information set h and some active player i at h. By our induction assumption, we have that $\Gamma^{k-1}(h) = D_i^{k-1} \times D_{-i}^{k-1}$ for some sets $D_i^{k-1} \subseteq S_i(h)$ and $D_{-i}^{k-1} \subseteq S_{-i}(h)$. We now prove that $\Gamma^k(h) = D_i^k \times D_{-i}^k$ for some sets $D_i^k \subseteq S_i(h)$ and $D_{-i}^k \subseteq S_{-i}(h)$.

Consider two arbitrary strategy combinations (s_i, s_{-i}) and (s_i', s_{-i}') in $\Gamma^k(h)$. We show that (s_i, s_{-i}') and (s_i', s_{-i}) are also in $\Gamma^k(h)$, which would prove the desired statement.

As $\Gamma^k(h) \subseteq \Gamma^{k-1}(h)$ we have in particular that (s_i, s_{-i}) and (s_i', s_{-i}') are in $\Gamma^{k-1}(h)$, so s_i and s_i' are in D_i^{k-1} and s_{-i} and s_{-i}' are in D_{-i}^{k-1}. Since

$$\Gamma^k(h) = \Gamma^{k-1}(h) \backslash \bigcup_{h' \succcurlyeq h} sd(\Gamma^{k-1}(h'))$$

it follows that s_i and s_i' are both not strictly dominated in any $\Gamma^{k-1}(h_i')$ where h_i' weakly follows h and i is active. Similarly, both s_{-i} and s_{-i}' do not contain any strategy for any player $j \neq i$ that is strictly dominated in some $\Gamma^{k-1}(h_j')$ where h_j' weakly follows h and j is active. But then, by definition of $\Gamma^k(h)$, it follows that (s_i, s_{-i}') and (s_i', s_{-i}) are also in $\Gamma^k(h)$.

So we have shown that (s_i, s_{-i}') and (s_i', s_{-i}) are in $\Gamma^k(h)$ whenever (s_i, s_{-i}) and (s_i', s_{-i}') are in $\Gamma^k(h)$. But then, we conclude that there must be some sets $D_i^k \subseteq S_i(h)$ and $D_{-i}^k \subseteq S_{-i}(h)$ such that $\Gamma^k(h) = D_i^k \times D_{-i}^k$. By induction on k, the proof is complete. □

9.3. Backward dominance procedure delivers nonempty output

We now prove Theorem 5.3, which states that the backward dominance procedure delivers at every information set a decision problem with nonempty strategy sets. Let $S_i^k(h)$ denote the set of player i strategies in the decision problem $\Gamma^k(h)$ produced by round k of the backward dominance procedure. We show, by induction on k, that $S_i^k(h)$ is always nonempty.

For $k = 0$ it is true since $S_i^0(h) = S_i(h)$, which is nonempty.

Suppose now that $k \geqslant 1$, and that $S_i^{k-1}(h)$ is nonempty for every information set h and player i. Fix some information set h^* and player i. We show that $S_i^k(h^*)$ is nonempty. By Lemma 9.3 we know that the collection $(S_{-i}^{k-1}(h_i))_{h_i \in H_i}$ satisfies the forward inclusion property. Hence, by Lemma 9.2, we can find a conditional belief vector $(b_i(h_i))_{h_i \in H_i}$ with $b_i(h_i) \in \Delta(S_{-i}^{k-1}(h_i))$ for all $h_i \in H_i$, and a strategy $s_i \in S_i(h^*)$, such that s_i is optimal at every $h_i \in H_i$ weakly following h^* for the belief $b_i(h_i)$. But then, we know from Lemma 9.4 that $s_i \in S_i^k(h^*)$, and hence $S_i^k(h^*)$ is nonempty. By induction on k, the proof is complete. □

9.4. Backward dominance procedure characterizes strategy choices under common belief in future rationality (without Bayesian updating)

We now prove our main result, Theorem 5.4, which states that the backward dominance procedure yields exactly those strategies that can rationally be chosen under common belief in future rationality, if we do not impose (common belief in) Bayesian updating. For the sake of brevity, whenever we speak about common belief in future rationality in this section, we actually mean common belief in future rationality without (common belief in) Bayesian updating.

So we must prove two directions: First, that every strategy that can rationally be chosen under common belief in future rationality survives the backward dominance procedure, and second that every strategy surviving the procedure can rationally be chosen under common belief in future rationality.

(a) Every strategy that can rationally be chosen under common belief in future rationality (without Bayesian updating) survives the backward dominance procedure.

For every player i and every information set $h_i \in H_i$, let

$$B_i(h_i) := \big\{ b_i(h_i) \in \Delta\big(S_{-i}(h_i)\big) \colon \text{there is a type } t_i \text{ expressing common belief in} \\ \text{future rationality such that the marginal of } b_i(t_i, h_i) \text{ on } S_{-i}(h_i) \text{ is } b_i(h_i) \big\}.$$

So, $B_i(h_i)$ contains those conditional beliefs at h_i about the opponents' strategy choices that are possible under common belief in future rationality. By $S_{-i}^k(h_i)$ we denote the set of opponents' strategies in the decision problem $\Gamma^k(h_i)$ produced in round k of the backward dominance procedure. We prove the following claim.

Claim. *$B_i(h_i) \subseteq \Delta(S_{-i}^k(h_i))$ for every k.*

Proof. We prove the claim by induction on k. For $k = 0$ the statement is true since $S_{-i}^0(h_i) = S_{-i}(h_i)$.

Now, take some $k \geqslant 1$, and assume that $B_i(h_i) \subseteq \Delta(S_{-i}^{k-1}(h_i))$ for every player i and every $h_i \in H_i$. Fix some player i and some information set $h_i \in H_i$. We show that $B_i(h_i) \subseteq \Delta(S_{-i}^k(h_i))$.

Take some $b_i(h_i) \in B_i(h_i)$. Then, there is some epistemic model $M = (T_i, b_i)_{i \in I}$, and some type $t_i \in T_i$ expressing common belief in future rationality, such that the marginal of $b_i(t_i, h_i)$ on $S_{-i}(h_i)$ is equal to $b_i(h_i)$. So, t_i's belief at h_i about the opponents' strategies and types, which is $b_i(t_i, h_i)$, only assigns positive probability to opponents' types t_j that express common belief in future rationality. Since, by our induction assumption, $B_j(h'_j) \subseteq \Delta(S_{-j}^{k-1}(h'_j))$ for all opponents j, and all $h'_j \in H_j$, it follows that $b_i(t_i, h_i)$ only assigns positive probability to opponents' types t_j whose belief at every $h'_j \in H_j$ about the other players' strategy choices is in $\Delta(S_{-j}^{k-1}(h'_j))$.

As t_i expresses common belief in future rationality, $b_i(t_i, h_i)$ only assigns positive probability to opponents' strategy–type pairs (s_j, t_j) where s_j is optimal for t_j at every $h'_j \in H_j(s_j)$ weakly following h_i. Together with the fact that $b_i(t_i, h_i)$ only assigns positive probability to opponents' types t_j whose belief at such h'_j about the other players' strategy choices is in $\Delta(S_{-j}^{k-1}(h'_j))$, this implies that $b_i(t_i, h_i)$ only assigns positive probability to opponents' strategies s_j that are optimal, at every $h'_j \in H_j(s_j)$ weakly following h_i, for some belief in $\Delta(S_{-j}^{k-1}(h'_j))$. However, by Lemma 9.4, these latter strategies s_j are exactly the strategies in $S_j^k(h_i)$. Hence, $b_i(t_i, h_i)$ only assigns positive probability to opponents' strategies in $S_j^k(h_i)$, which means that the marginal of $b_i(t_i, h_i)$ on $S_{-i}(h_i)$ is in $\Delta(S_{-i}^k(h_i))$. By definition, the marginal of $b_i(t_i, h_i)$ on $S_{-i}(h_i)$ was $b_i(h_i)$, so $b_i(h_i) \in \Delta(S_{-i}^k(h_i))$.

Since this holds for every $b_i(h_i) \in B_i(h_i)$, we may conclude that $B_i(h_i) \subseteq \Delta(S_{-i}^k(h_i))$. By induction on k, the proof of the claim is complete. □

We are now ready to prove part (a). Take some strategy s_i that can rationally be chosen under common belief in future rationality. Then, there is some epistemic model $M = (T_i, b_i)_{i \in I}$, and some type $t_i \in T_i$ expressing common belief in future rationality, such that s_i is rational for t_i. So, s_i must be optimal at every $h_i \in H_i(s_i)$ for the belief $b_i(t_i, h_i)$. By the claim above we know that $b_i(t_i, h_i) \in \Delta(S_{-i}^\infty(h_i))$, where $S_{-i}^\infty(h_i) := \bigcap_k S_{-i}^k(h_i)$. So, at every $h_i \in H_i(s_i)$ strategy s_i is optimal for some belief in $\Delta(S_{-i}^\infty(h_i))$. By Lemma 9.4 this implies that $s_i \in S_i^\infty(\emptyset)$, where $S_i^\infty(\emptyset) := \bigcap_k S_i^k(\emptyset)$. This means, however, that s_i survives the backward dominance procedure, and hence the proof of part (a) is complete.

(b) Every strategy that survives the backward dominance procedure can rationally be chosen under common belief in future rationality (without Bayesian updating).

For every information set h and every player i, let $S_i^\infty(h)$ be the set of player i strategies that are left at h at the end of the backward dominance procedure. So, $S_i^\infty(h) := \bigcap_k S_i^k(h)$. Remember that $S_i^\infty(\emptyset)$ contains exactly those player i strategies that survive the backward dominance procedure.

The idea for proving (b) is as follows: We construct an epistemic model $M = (T_i, b_i)_{i \in I}$ in which every type expresses common belief in future rationality. Moreover, for every $s_i \in S_i^\infty(\emptyset)$ there will be some type $t_i \in T_i$ for which s_i is rational. But then, every $s_i \in S_i^\infty(\emptyset)$ can be chosen rationally by a type that expresses common belief in future rationality, which would prove part (b).

For every player i, we define the set of types

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

For every strategy s_i, let $H_i^*(s_i)$ be the (possibly empty) collection of information sets $h_i \in H_i$ for which $s_i \in S_i^\infty(h_i)$. So, by Lemma 9.4, we can find for every $s_i \in S_i$ some conditional belief vector $(b_i(s_i, h_i))_{h_i \in H_i}$ such that (a) $b_i(s_i, h_i) \in \Delta(S_{-i}^\infty(h_i))$ for every $h_i \in H_i$, and (b) s_i is optimal at every $h_i \in H_i^*(s_i)$ for the belief $b_i(s_i, h_i)$.

We will now define the conditional beliefs of the types. Take a type $t_i^{s_i}$ in T_i, and an information set $h_i \in H_i$. For every opponents' strategy profile $(s_j)_{j \neq i}$, let $b_i(s_i, h_i)((s_j)_{j \neq i})$ be the probability that $b_i(s_i, h_i)$ assigns to $(s_j)_{j \neq i}$. Let $b_i(t_i^{s_i}, h_i)$ be the conditional belief about the opponents' strategy–type pairs given by

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

So, at every $h_i \in H_i$, type $t_i^{s_i}$ holds the same belief about the opponents' strategy choices as $b_i(s_i, h_i)$. Moreover, at every information set $h_i \in H_i$, type $t_i^{s_i}$ assigns only positive probability to strategy–type pairs (s_j, t_j) where $s_j \in S_j^\infty(h_i)$ and $t_j = t_j^{s_j}$.

We now prove that every type in this epistemic model believes in the opponents' future rationality. Take some type $t_i^{s_i} \in T_i$ and an information set $h \in H_i$. Then, by construction, $b_i(t_i^{s_i}, h_i)$ only assigns positive probability to opponents' strategy–type pairs $(s_j, t_j^{s_j})$ where $s_j \in S_j^\infty(h_i)$.

Take an opponent's strategy $s_j \in S_j^\infty(h_i)$. By construction of our algorithm, we have that $s_j \in S_j^\infty(h'_j)$ for every $h'_j \in H_j(s_j)$ weakly following h_i. In other words, if $s_j \in S_j^\infty(h_i)$, then every $h'_j \in H_j(s_j)$ weakly following h_i is in $H_j^*(s_j)$.

By construction, at every $h'_j \in H_j^*(s_j)$ type $t_j^{s_j}$ holds the same belief about the opponents' strategy choices as $b_j(s_j, h'_j)$. Moreover, at every $h'_j \in H_j^*(s_j)$, strategy s_j is optimal under the belief $b_j(s_j, h'_j)$. So, at every $h'_j \in H_j^*(s_j)$, strategy s_j is optimal for type $t_j^{s_j}$. Since we have seen that every $h'_j \in H_j(s_j)$ weakly following h_i is in $H_j^*(s_j)$, it follows that s_j is optimal for type $t_j^{s_j}$ at every $h'_j \in H_j(s_j)$ that weakly follows h_i.

So, we have shown for every $s_j \in S_j^\infty(h_i)$ that s_j is optimal for type $t_j^{s_j}$ at every $h'_j \in H_j(s_j)$ weakly following h_i. Since $b_i(t_i^{s_i}, h_i)$ only assigns positive probability to opponents' strategy–type pairs $(s_j, t_j^{s_j})$ where $s_j \in S_j^\infty(h_i)$, we may conclude the following: Type $t_i^{s_i}$ assigns at h_i only positive probability to opponents' strategy–type pairs $(s_j, t_j^{s_j})$ where s_j is optimal for type $t_j^{s_j}$ at every $h'_j \in H_j(s_j)$ weakly following h_i. In other words, type $t_i^{s_i}$ believes at h_i in the opponents' future rationality. As this applies to every $h_i \in H_i$, we may conclude that type $t_i^{s_i}$ believes in the opponents' future rationality. So, every type $t_i^{s_i}$ in the epistemic model believes in the opponents' future rationality.

From this fact, it immediately follows that every type in the epistemic model expresses common belief in future rationality.

Now, take a strategy s_i that survives the backward dominance procedure, that is, $s_i \in S_i^\infty(\emptyset)$. Consider the associated type $t_i^{s_i}$. Above, we have seen that every $h_i \in H_i(s_i)$ weakly following $\emptyset$ is in $H_i^*(s_i)$. Since, as we have seen above, s_i is optimal for $t_i^{s_i}$ at every $h_i \in H_i^*(s_i)$, it follows that s_i is optimal for $t_i^{s_i}$ at every $h_i \in H_i(s_i)$ weakly following $\emptyset$. However, this means that s_i is rational for type $t_i^{s_i}$. Since, as we have shown above, $t_i^{s_i}$ expresses common belief in future rationality, it follows that s_i can rationally be chosen under common belief in future rationality. This completes the proof of part (b). □

9.5. Best-response characterization

We finally prove Theorem 6.3, which provides a best-response characterization of common belief in future rationality. More precisely, we must show that a strategy s_i can rationally be chosen under common belief in future rationality, if and only if, there is a collection $(D_i(h))_{h \in H, i \in I}$ of strategy subsets which is closed under belief in future rationality and where $s_i \in D_i(\emptyset)$. So, we must prove two directions.

Suppose first that s_i can rationally be chosen under common belief in future rationality. Then, by Theorem 4.3, $s_i \in S_i^\infty(\emptyset)$. As the collection $(S_i^\infty(h))_{h\in H, i\in I}$ of strategy subsets is closed under belief in future rationality, the proof of the first direction is complete.

Suppose next that $(D_i(h))_{h\in H, i\in I}$ is a collection of strategy subsets which is closed under belief in future rationality, and take some $s_i \in D_i(\emptyset)$. We must show that s_i can rationally be chosen under common belief in future rationality. To show this we prove the following claim. Recall that $S_i^k(h)$ denotes the set of player i strategies produced in round k of the backwards rationalizability procedure at information set h.

Claim. *$D_i(h) \subseteq S_i^k(h)$ for every k.*

Proof. We proceed by induction on k. For $k = 0$ the statement is true since $S_i^0(h) = S_i(h)$.

Take now some $k \geqslant 1$, and suppose that $D_i(h) \subseteq S_i^{k-1}(h)$ for every player i and information set h. Fix some player i and some information set h. We will show that $D_i(h) \subseteq S_i^k(h)$.

Choose some arbitrary $s_i \in D_i(h)$. As the collection $(D_i(h))_{h\in H, i\in I}$ is closed under belief in future rationality, there must be some conditional belief vector b_i satisfying Bayesian updating such that (1) $b_i(h_i') \in \Delta(D_{-i}(h_i'))$ for every $h_i' \in H_i$ weakly following h, and (2) s_i is optimal for $b_i(h_i')$ at every $h_i' \in H_i(s_i)$ weakly following h. As, by induction assumption, $D_{-i}(h_i') \subseteq S_{-i}^{k-1}(h_i')$, we conclude that there is some conditional belief vector b_i satisfying Bayesian updating such that (1) $b_i(h_i') \in \Delta(S_{-i}^{k-1}(h_i'))$ for every $h_i' \in H_i$ weakly following h, and (2) s_i is optimal for $b_i(h_i')$ at every $h_i' \in H_i(s_i)$ weakly following h. But then, by construction of the backwards rationalizability procedure, we have that $s_i \in S_i^k(h)$. We thus conclude that $D_i(h) \subseteq S_i^k(h)$, and the proof of the claim is complete by induction on k. □

From the claim, it immediately follows that $D_i(h) \subseteq S_i^\infty(h)$ for every information set h and player i. Take some strategy $s_i \in D_i(\emptyset)$. As $D_i(\emptyset) \subseteq S_i^\infty(\emptyset)$, it follows that $s_i \in S_i^\infty(\emptyset)$, which means that s_i survives the backwards rationalizability procedure. But then, by Theorem 4.3, we know that s_i can rationally be chosen under common belief in future rationality. This completes the proof of Theorem 6.3. □

References

Asheim, G.B., 2002. On the epistemic foundation for backward induction. Math. Soc. Sci. 44, 121–144.

Asheim, G.B., Perea, A., 2005. Sequential and quasi-perfect rationalizability in extensive games. Games Econ. Behav. 53, 15–42.

Baltag, A., Smets, S., Zvesper, J.A., 2009. Keep 'hoping' for rationality: a solution to the backward induction paradox. Synthese 169, 301–333. Knowledge, Rationality and Action, pp. 705–737.

Battigalli, P., 1997. On rationalizability in extensive games. J. Econ. Theory 74, 40–61.

Battigalli, P., Siniscalchi, M., 2002. Strong belief and forward induction reasoning. J. Econ. Theory 106, 356–391.

Bernheim, B.D., 1984. Rationalizable strategic behavior. Econometrica 52, 1007–1028.

Chen, J., Micali, S., 2013. The order independence of iterated dominance in extensive games. Theoretical Econ. 8, 125–163.

Dekel, E., Fudenberg, D., Levine, D.K., 1999. Payoff information and self-confirming equilibrium. J. Econ. Theory 89, 165–185.

Dekel, E., Fudenberg, D., Levine, D.K., 2002. Subjective uncertainty over behavior strategies: A correction. J. Econ. Theory 104, 473–478.

Elmes, S., Reny, P.J., 1994. On the strategic equivalence of extensive form games. J. Econ. Theory 62, 1–23.

Feinberg, Y., 2005. Subjective reasoning–dynamic games. Games Econ. Behav. 52, 54–93.

Hendon, E., Jacobsen, H.J., Sloth, B., 1996. The one-shot-deviation principle for sequential rationality. Games Econ. Behav. 12, 274–282.

Kreps, D.M., Wilson, R., 1982. Sequential equilibria. Econometrica 50, 863–894.

Pearce, D.G., 1984. Rationalizable strategic behavior and the problem of perfection. Econometrica 52, 1029–1050.

Penta, A., 2009. Robust dynamic mechanism design. Manuscript. University of Pennsylvania.

Perea, A., 2001. Rationality in Extensive Form Games. Theory Decis. Libr., Ser. C Game Theory Math. Program. Oper. Res. Kluwer Academic Publishers, Boston/Dordrecht/London.

Perea, A., 2002. A note on the one-deviation property in extensive games. Games Econ. Behav. 40, 322–338.

Perea, A., 2007. Epistemic foundations for backward induction: an overview. In: van Benthem, Johan, Gabbay, Dov, Löwe, Benedikt (Eds.), Interactive Logic Proceedings of the 7th Augustus de Morgan Workshop, London. In: Texts in Logic and Games, vol. 1. Amsterdam University Press, pp. 159–193.

Perea, A., 2010. Backward induction versus forward induction reasoning. Games 1, 168–188.

Perea, A., 2012. Epistemic Game Theory: Reasoning and Choice. Cambridge University Press.

Rubinstein, A., 1991. Comments on the interpretation of game theory. Econometrica 59, 909–924.

Samet, D., 1996. Hypothetical knowledge and games with perfect information. Games Econ. Behav. 17, 230–251.

Shimoji, M., Watson, J., 1998. Conditional dominance, rationalizability, and game forms. J. Econ. Theory 83, 161–195.

Tan, T., Werlang, S.R.C., 1988. The Bayesian foundations of solution concepts of games. J. Econ. Theory 45, 370–391.

Thompson, F.B., 1952. Equivalence of games in extensive form. Discussion paper RM 759. The RAND Corporation.