Belief in the Opponents' Future Rationality

Andrés Perea
Maastricht University

This version: January 2010

Abstract

For dynamic games we consider the idea that a player, at every stage of the game, believes that his opponents will choose rationally in the future. Not only this, we also assume that players, throughout the game, believe that their opponents always believe that their opponents will choose rationally in the future, and so on. This leads to the concept of *common belief in future rationality,* which we formalize within an epistemic model. Our main contribution is to present an easy elimination procedure, *backwards dominance,* that selects exactly those strategies that can rationally be chosen under common belief in future rationality. The algorithm proceeds by successively eliminating strategies at every information set of the game. More specifically, in round k of the procedure we eliminate at a given information set h those strategies for player i that are strictly dominated at some player i information set h' *weakly following* h, given the opponents' strategies that have survived at h' until round k.

Key words and phrases: Epistemic game theory, dynamic games, belief in future rationality, algorithms.

Journal of Economic Literature Classification: C72

Contact information:
Maastricht University, Department of Quantitative Economics
P.O. Box 616, 6200 MD Maastricht, The Netherlands
a.perea@maastrichtuniversity.nl
Webpage: http://www.personeel.unimaas.nl/a.perea/

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 major difficulty 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 phrase. 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.

We do not only impose, however, 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 dynamic games, 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.

Our main contribution is that we deliver an easy algorithm, called *backwards dominance,* which generates for every player exactly those strategies he can rationally choose under common

[1] At least, if we do not allow player 2 to revise his belief about player 1's utility function. Perea (2006, 2007a) considers a model where a player may revise his belief about an opponent's utility function upon observing an unexpected move. Within this framework, it is possible to require that a player always believes that his opponents have chosen rationally in the past.

belief in future rationality. The algorithm proceeds by successively eliminating, at every information set, some strategies for the players. 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 for player i, given the opponents' strategies that have survived until round k at that information set h. We continue until we cannot eliminate anything more. The strategies that eventually survive at the beginning of the game are those that survive the algorithm. The main result in this paper shows that the strategies that survive the backwards dominance procedure are exactly the strategies that can rationally be chosen under common belief in future rationality.

Some important properties of the algorithm are that it always stops after finitely many steps, that it always delivers a nonempty set of strategies for every player, and that the order and speed in which we eliminate strategies from the game does not matter for the eventual result. The second of these properties, together with our main theorem, implies that common belief in future rationality is always possible in every game, that is, it never leads to logical contradictions.

If we apply the backwards dominance procedure to games with perfect information, then we obtain precisely the well-known backwards induction procedure. As a consequence, applying common belief in future rationality to games with perfect information leads to backwards 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 backwards induction. See, for instance, Asheim (2002), Feinberg (2005) and Samet (1996). The reader may consult Perea (2007b) for a detailed overview of the various epistemic foundations that have been offered for backwards induction in the literature. For general dynamic games, belief in the opponents' future rationality is *implicitly* present in "backwards induction concepts" such as sequential equilibrium (Kreps and Wilson (1982)) and sequential rationalizability (Dekel, Fudenberg and Levine (1999, 2002) and Asheim and Perea (2005)). In fact, we show in Section 6 that sequential equilibrium and sequential rationalizability are both more restrictive than common belief in future rationality. Moreover, the difference with sequential rationalizability only lies in the fact that the latter assumes (common belief in) Kreps-Wilson consistency of beliefs, and independent beliefs about the opponents' future choices, whereas common belief in future rationality does not. Independently from our paper, Penta (2009) has developed a procedure, *backwards rationalizability,* which is similar to our backwards dominance procedure. A difference with our procedure is that backwards rationalizability requires (common belief in) Bayesian updating, whereas we do not. Also, Penta's procedure works by successively eliminating conditional belief vectors and strategies from the game, whereas our procedure only works on strategies.

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 are imposed. In particular, we impose no equilibrium conditions as in sequential equilibrium. 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. In the game of Figure 1, for instance, sequential equilibrium rules out strategy (a, c) for player 1, but we argue in Section 4.2 that this strategy is perfectly reasonable for player 1. Common belief in future rationality does not rule out this strategy. Finally, we provide an easy algorithm that supports 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 6 we also compare our notion with the concept of *extensive form rationalizability* (Pearce (1984)) 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. Despite this, we can nicely compare our backwards dominance procedure with the *iterated conditional dominance procedure* (Shimoji and Watson (1998)), which leads to 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 6 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 introduce the backwards dominance algorithm, illustrate it by means of an example, and present the main theorem stating that the algorithm selects exactly those strategies that can rationally be chosen under common belief in future rationality. In Section 5 we discuss some important properties of the algorithm, and use these to derive some additional insights about the concept of common belief in future rationality. In Section 6 we explore the relation between common belief in future rationality, and other concepts for dynamic games such as sequential rationalizability, backwards rationalizability and extensive form rationalizability. In Section 7 we discuss possible lines for future research. Section 8 contains all the proofs.

2. Model

In this section we lay out some basic concepts, such as the notion of a dynamic game and the concept of a strategy, and explain how to build an epistemic model for such dynamic games.

2.1. Dynamic Games

The model of a dynamic game presented here is a bit more general than usual, as we explicitly allow for *simultaneous* choices by players at certain stages of the game. In our setup, a *dynamic game* is a tuple $\Gamma = (I, X, Z, (X_i, C_i, H_i, u_i)_{i \in I})$ where

(a) I is the set of *players;*

(b) X is the set of *non-terminal histories.* Every non-terminal history $x \in X$ represents a situation where one or more players must make a choice;

(c) Z is the set of *terminal histories.* Every terminal history $z \in Z$ represents a situation where the game ends;

(d) $X_i \subseteq X$ is the set of histories at which player i must make a choice. At every history $x \in X$ at least one player must make a choice, that is, for every $x \in X$ there is at least some i with $x \in X_i$. However, for a given history x there may be various players i with $x \in X_i$. This models a situation where various players simultaneously choose at x. For a given history $x \in X$, we denote by $I(x) := \{i \in I : x \in X_i\}$ the set of *active players* at x;

(e) C_i assigns to every history $x \in X_i$ the set of *choices* $C_i(x)$ from which player i can choose at x;

(f) H_i is the collection of *information sets* for player i. Formally, $H_i = \{h_i^1, ..., h_i^K\}$ where $h_i^k \subseteq X_i$ for every k, the sets h_i^k are mutually disjoint, and $X_i = \cup_k h_i^k$. The interpretation of an information set $h \in H_i$ is that at h player i knows that some history in h has been realized, without knowing precisely which one;

(g) u_i is player i's *utility function,* assigning to every terminal history $z \in Z$ some utility $u_i(z)$ in $\Re$.

Throughout this paper we assume that all sets above are finite. The histories in X and Z consist of finite sequences of choice combinations

$$((c_i^1)_{i \in I^1}, (c_i^2)_{i \in I^2}, ..., (c_i^K)_{i \in I^K}),$$

where $I^1, ..., I^K$ are nonempty subsets of players, such that

(a) $\emptyset$ (the empty sequence) is in X,

(b) if $x \in X$ and $(c_i)_{i \in I(x)} \in \prod_{i \in I(x)} C_i(x)$, then $(x, (c_i)_{i \in I(x)}) \in X \cup Z$,

(c) if $z \in Z$, then there is no choice combination $(c_i)_{i \in \hat{I}}$ such that $(z, (c_i)_{i \in \hat{I}}) \in X \cup Z$,

(d) for every $x \in X \cup Z$, $x \neq \emptyset$, there is a unique $y \in X$ and $(c_i)_{i \in I(y)} \in \prod_{i \in I(y)} C_i(y)$ such that $x = (y, (c_i)_{i \in I(y)})$.

Hence, a history $x \in X \cup Z$ represents the sequence of choice combinations that have been made by the players until this moment.

Moreover, we assume that the collections H_i of information sets are such that

(a) two histories in the same information set for player i have the same set of available choices for player i. That is, for every $h \in H_i$, and every $x, y \in h$, it holds that $C_i(x) = C_i(y)$. This condition must hold since player i is assumed to know his set of available choices at h. We can thus speak of $C_i(h)$ for a given information set $h \in H_i$;

(b) two histories in the same information set for player i must pass through exactly the same collection of information sets for player i, and must hold exactly the same past choices for player i. This condition guarantees that player i has *perfect recall,* that is, at every information set $h \in H_i$ player i remembers the information he possessed before, and the choices he made before.

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 history $x = (y, (c_i^1)_{i \in I^1}, (c_i^2)_{i \in I^2}, ..., (c_i^K)_{i \in I^K})$ for some choice-combinations $(c_i^1)_{i \in I^1}, (c_i^2)_{i \in I^2}, ..., (c_i^K)_{i \in I^K}$. The information sets h and h' are called *simultaneous* if there is some history x with $x \in h$ and $x \in h'$. Finally, we say that information set h *weakly follows* h' if either h follows h', or h and h' are simultaneous.

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 strategy. Formally, for every $h, h' \in H_i$ such that h precedes h', let $c_i(h, h')$ be the choice at h for player i that leads to h'. Note that $c_i(h, h')$ 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 $h \in \hat{H}_i$ some choice $s_i(h) \in C_i(h)$. We say that s_i *possibly reaches* an information set h if at every $h' \in H_i$ preceding h we have that $s_i(h') = c_i(h', h)$. By $H_i(s_i)$ we denote the collection of player i information sets that s_i possibly reaches. A *strategy* for player i is a function s_i, assigning to every $h \in \hat{H}_i \subseteq H_i$ some choice $s_i(h) \in C_i(h)$, such that $\hat{H}_i = H_i(s_i)$. For a given information set h, denote by $S_i(h)$ the set of strategies for player i that possibly reach h. By $S_{-i}(h)$ we denote the strategy profiles for i's opponents that possibly reach h, that is, $s_{-i} \in S_{-i}(h)$ if there is some $s_i \in S_i(h)$ such that (s_i, s_{-i}) reaches some history in h. By $S(h)$ we denote the set of strategy profiles $(s_i)_{i \in I}$ that reach some history in h. By perfect recall we have that $S(h) = S_i(h) \times S_{-i}(h)$ for every player i and every information set $h \in H_i$.

2.3. Epistemic Model

We now wish to model the players' beliefs in the game. At every information set $h \in H_i$, 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 2.1. *(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* $h \in H_i$*, a probability distribution* $b_i(t_i, h) \in \Delta(S_{-i}(h) \times T_{-i})$*.*

Recall that $S_{-i}(h)$ represents the set of opponents' strategy combinations that possibly reach h. 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 $h \in H_i$ type t_i holds a conditional probabilistic belief $b_i(t_i, h)$ 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 $h \in H_i$ 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.

3. Belief in the Opponents' Future Rationality

We now present the main idea in this paper, namely that a player should always believe 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 a history $h \in H_i(s_i)$ that is possibly reached by s_i. By $u_i(s_i, t_i \mid h)$ we denote the expected utility from choosing s_i under the conditional belief that t_i holds at h 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* $h \in H_i(s_i)$*. Strategy* s_i *is optimal for type* t_i *at* h *if* $u_i(s_i, t_i \mid h) \geq u_i(s_i', t_i \mid h)$ *for all* $s_i' \in S_i(h)$*.*

Remember that $S_i(h)$ is the set of player i strategies that possibly reach h. 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* $h \in H_i$*, and an opponent* $j \neq i$*. Type* t_i *believes at* h *in* j*'s future rationality if* $b_i(t_i, h)$ *only assigns positive probability to* j*'s strategy-type pairs* (s_j, t_j) *where* s_j *is optimal for* t_j *at every* $h' \in H_j(s_j)$ *that weakly follows* h*. Type* t_i *believes in the opponents' future rationality if at every* $h \in H_i$*, 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.

Next, we formalize the requirement that a player should not only believe in the opponents' future rationality, but should also always believe that every opponent believes in his opponents' future rationality, and so on.

Definition 3.3. *(Common belief in future rationality) Type t_i expresses common belief in future rationality if (a) t_i believes in the opponents' future rationality, (b) t_i assigns, at every information set, only positive probability to opponents' types that believe in their opponents' future rationality, (c) t_i assigns, at every information set, only positive probability to opponents' types that, at every information set, only assign positive probability to opponents' types that believe in the opponents' future rationality, and so on.*

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

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

Definition 3.5. *(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 belief hierarchy expressing common belief in future rationality that supports this strategy choice.

4. Algorithm

In this section we present a procedure, called *backwards dominance,* that iteratedly eliminates strategies from the game. We prove that this procedure generates exactly those strategies that can rationally be chosen under common belief in future rationality.

4.1. Description of the Algorithm

In order to formally state our algorithm we need the following definitions. Consider an information set $h \in H_i$ for player i, a subset $D_i \subseteq S_i(h)$ of strategies for player i that possibly reach h, and a subset $D_{-i} \subseteq S_{-i}(h)$ of strategy combinations for i's opponents possibly reaching h. Then, (D_i, D_{-i}) is called a *conditional game* for player i at h, and we say that player i is *active* at this conditional game. Note that several players may be active at the same conditional game, since several players may make a simultaneous move at the associated information set. Within a conditional game (D_i, D_{-i}) for player i, a strategy $s_i \in D_i$ is called *strictly dominated* 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 $s_{-i} \in D_{-i}$. A conditional game at h is said to weakly follow an information set h' if h weakly follows h'. For a given information set $h \in H_i$, the *full* conditional game at h is the conditional game $(S_i(h), S_{-i}(h))$ where no strategies have been eliminated yet.

Algorithm 4.1. *(Backwards dominance procedure)*

Initial step. *For every information set h, let $\Gamma^0(h)$ be the full conditional game at h.*

Inductive step. *Let $k \geq 1$, and suppose that the conditional games $\Gamma^{k-1}(h)$ have been defined for every information set h. Then, at every information set h delete from the conditional game $\Gamma^{k-1}(h)$ those strategies s_i for player i that are strictly dominated within some conditional game $\Gamma^{k-1}(h')$ for player i that weakly follows h. This yields the new conditional games $\Gamma^k(h)$. Continue this procedure until no further strategies can be eliminated in this way.*

Since we only have finitely many strategies in the game, and the conditional games can only become smaller at every step, this procedure must converge after finitely many steps. An important question though 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 backwards dominance procedure at some information set h if s_i is part of the conditional game $\Gamma^k(h)$ for all k.

Theorem 4.2. *(Algorithm 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 backwards dominance procedure at h.*

The formal proof for this result can be found in Section 8.

4.2. Illustration of the Algorithm

In this section we will illustrate our backwards dominance procedure by means of an example. Consider the game in Figure 1. So, at the beginning of the game, $\emptyset$, only player 1 is active. He

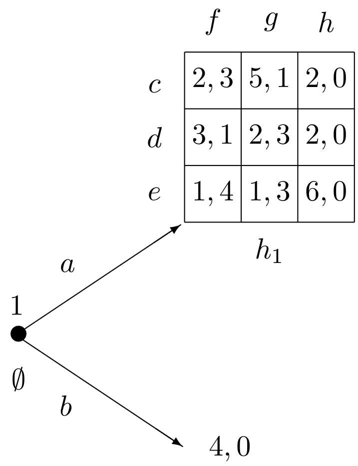

Figure 1: Illustration of the backwards dominance procedure

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.

At the beginning of the procedure we start with two conditional games, namely the full conditional game $\Gamma^0(\emptyset)$ at $\emptyset$ where only player 1 is active, and the full conditional game $\Gamma^0(h_1)$ at h_1 where both players are active. These conditional games can be found in Table 1.

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

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 1: Full conditional games in backwards dominance procedure

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

Pl. 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 2: Step 1 of backwards dominance procedure

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

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

Table 3: Step 2 of backwards dominance procedure

at the future conditional game $\Gamma^0(h_1)$ 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 conditional games $\Gamma^1(\emptyset)$ and $\Gamma^1(h_1)$ which can be found in Table 2.

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 conditional games $\Gamma^2(\emptyset)$ and $\Gamma^2(h_1)$ presented in Table 3.

After this step no more strategies can be eliminated. So, the algorithm stops here. Note that at the beginning of the game, the strategies (a,c) and b have survived for player 1, and the strategies f and g have survived for player 2. Our Theorem 4.3 below states that these are exactly the strategies that can rationally be chosen 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 $\frac{1}{4}$ we 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. Main Result

We now show that the backwards dominance procedure generates exactly those strategies that can rationally be chosen under common belief in future rationality. Say that a strategy s_i survives the backwards dominance procedure if s_i is in $\Gamma^k(\emptyset)$ for every k.

Theorem 4.3. *(Algorithm works) A strategy s_i can rationally be chosen under common belief in future rationality if and only if s_i survives the backwards dominance procedure.*

The proof can be found in Section 8. In particular, our theorem shows that common belief in future rationality is always possible for every player: Take, namely, an arbitrary player i in the game. Then, we know from Theorem 4.2 that there is at least one strategy s_i for this player that survives backwards dominance. Theorem 4.3 then guarantees that for this strategy s_i we can find an epistemic model, and a type t_i for player i within it, such that t_i expresses common belief in future rationality, and such that s_i is rational for t_i. In particular, we can always construct for every player some type that expresses common belief in future rationality.

5. Discussion of the Algorithm

In this section we will discuss some important properties of our backwards dominance procedure, and use these to derive some new insights about common belief in future rationality.

5.1. Order Independence

As we defined it, the backwards dominance algorithm eliminates, at every step and every information set h, *all* strategies for player i that are strictly dominated at some conditional game 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 backwards 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 much 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 conditional game $\tilde{\Gamma}^1(h')$ for player i. Here, $\tilde{\Gamma}^1(h')$ is the conditional game at h' after Step 1 of Procedure 2. Now, let $\Gamma^1(h')$ be the conditional game at h' after Step 1 of Procedure 1. Then, $\Gamma^1(h')$ contains at most as much 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 much 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 conditional game $\Gamma^{k-1}(h')$ for player i weakly following h. By assumption, in Procedure 2 there is some step $m \geq k-1$ such that the associated conditional game $\tilde{\Gamma}^m(h')$ 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 backwards dominance procedure. The intuitive reason is that the algorithm is *monotonic* in the following sense: If we make the conditional games smaller, then it becomes easier for a strategy to become strictly dominated, and hence we will eliminate more, which in turn leads to smaller conditional games, and so on.

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

5.2. 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 history exactly one player is active, and this player knows exactly which choices have been made until then.

Formally, this means that at every history x there is exactly one player i with $x \in X_i$, and the singleton $\{x\}$ is an information set for player i.

Say that a game with perfect information is *generic* if for every player i, and every history $x \in X_i$, two different choices at x will always lead to two different utilities for player i. That is, for every two terminal histories z, z' following $x \in X_i$ which contain different choices at x, we have that $u_i(z) \neq u_i(z')$. It is easily seen that every generic game with perfect information yields a unique backwards induction strategy for every player.

Consider now an arbitrary generic game with perfect information. We know that the backwards dominance procedure delivers exactly the strategies that can rationally be made 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 backwards induction order described above. So, we first consider all conditional games at the end of the game, and apply the backwards dominance procedure there. This, however, amounts to deleting all suboptimal choices at each of the information sets at the end of the game. That is, we uniquely select the backwards induction choices at all information sets at the end of the game.

Next, we turn to the conditional games just before these, and apply our backwards dominance procedure there. This means that at these information sets we first delete the strategies that were already deleted at the previous round. In this case, we would thus delete all strategies that would not prescribe the backwards induction choice at the last information sets in the game. So, we would keep only those strategies that *do* prescribe the backwards induction choices 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 backwards induction choices at the end of the game. Hence, we keep only those strategies that prescribe choices that are optimal against the backwards induction choices at the end of the game. So, we select the backwards 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 dominance procedure in the backwards induction fashion would yield exactly the backwards induction choice at every information set. Consequently, we obtain the unique backwards induction strategy for every player. Since the order of elimination does not matter, as we have seen, we conclude that applying the backwards dominance procedure to a generic game with perfect information would yield precisely the backwards induction strategies for the players.

Together with our Theorem 4.3 we thus see that in every generic game with perfect information, common belief in future rationality leads to backwards induction.

Theorem 5.1. *(Common belief in future rationality leads to backwards induction) Consider a generic dynamic game with perfect information. Then, every player has exactly one strategy he can rationally choose under common belief in future rationality, namely his backwards induction strategy.*

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

5.3. Best-Response Characterization

We will finally use the algorithm 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 dominance procedure at h. By construction of the algorithm, these sets $S_i^\infty(h)$ have the following property: If $s_i \in S_i^\infty(h)$, then at every $h' \in H_i(s_i)$ weakly following h strategy s_i is not strictly dominated on $S_{-i}^\infty(h')$. Here, $S_{-i}^\infty(h') := \prod_{j \neq i} S_j^\infty(h')$.

By Lemma 3 in Pearce (1984) we know that s_i is not strictly dominated on $S_{-i}^\infty(h')$ if and only if s_i is optimal at h' for some belief $b_i(h') \in \Delta(S_{-i}^\infty(h'))$. So, if $s_i \in S_i^\infty(h)$, then at every $h' \in H_i(s_i)$ weakly following h there is some belief $b_i(h') \in \Delta(S_{-i}^\infty(h'))$ for which s_i is optimal at 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.

Definition 5.2. *(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 $s_i \in D_i(h)$, and every $h' \in H_i(s_i)$ weakly following h, there is some belief $b_i(h') \in \Delta(D_{-i}(h'))$ for which s_i is optimal.*

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

Theorem 5.3. *(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 8. 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 dominance 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.

6. 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, backwards rationalizability and extensive form rationalizability.

6.1. Sequential Rationalizability

The concept of *sequential rationalizability* has been proposed independently by Dekel, Fudenberg and Levine (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 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 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.

Note that 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 $(a_i^n, \pi^{i,n}_{-i})_{n\in\mathbb{N}}$ converging to (a_i, π^i_{-i}) in which $\pi^{i,n}_{-i}$ assigns positive probability to all choices, and a_i^n is obtained from $\pi^{i,n}_{-i}$ 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 $h \in H_i$ 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 are part of some triple in V.[2]

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 *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 $(b_i(h))_{h\in H_i}$ in our setup, where $b_i(h)$ is a probability distribution over $S_{-i}(h)$ for every $h \in H_i$. This conditional belief vector $(b_i(h))_{h\in H_i}$ should be such, however, that i's conditional belief at h about the opponents' future behavior is independent across opponents. For every player i, consider a set $\tilde{V}_i$ of conditional belief vectors $(b_i(h))_{h\in H_i}$. Then, the collection $\tilde{V} = (\tilde{V}_i)_{i\in I}$ is *sequentially rationalizable* if for every $(b_i(h))_{h\in H_i} \in \tilde{V}_i$,

[2] For a precise statement of this condition, see Definition 2.2 in Dekel, Fudenberg and Levine (2002).

(d) at every h, the conditional belief about the opponents' future behavior is independent across opponents,

(e) the conditional belief vector $(b_i(h))_{h \in H_i}$ is Kreps-Wilson consistent,

(f) at every $h \in H_i$, the conditional belief $b_i(h)$ only assigns positive probability to opponents' strategies s_j which, at every $h' \in H_j(s_j)$ *weakly following* h, are optimal 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 rational collection $(\tilde{V}_i)_{i \in I}$ of conditional belief vectors. For every player i, every information set $h \in H_i$, and every opponent j, let $D_j(h) \subseteq S_j(h)$ be the set of strategies that receive positive probability at h under some conditional belief in $\tilde{V}_i$. At an information set $h \in H_i$, let $D_i(h) \subseteq S_i(h)$ be the set of strategies in $S_i(h)$ that, at every $h' \in H_i$ weakly following h, are optimal for some belief in $\Delta(D_{-i}(h'))$. By condition (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 at every $h' \in H_i(s_i)$ weakly following h there is some $b_i(h') \in \Delta(D_{-i}(h'))$ for which s_i is optimal. That is, the collection $(D_i(h))_{i \in I, h \in H}$ is closed under belief in future rationality, conform our Definition 5.2. We have thus shown that every sequentially rational 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 5.3 that every sequentially rational strategy can rationally be chosen under common belief in future rationality. We have thus established the following result.

Theorem 6.1. *(Relation to sequential rationalizability) Every sequentially rationalizable strategy can rationally be chosen under common belief in future rationality.*

It can be shown that the converse is not true: Not every strategy that can rationally be chosen under common belief in future rationality is sequentially rationalizable. Consider, for instance, the game in Figure 2. At the beginning of the game, $\emptyset$, player 1 chooses between a and b, and player 2 simultaneously chooses between c and d. 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_1 or information set h_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_3 after choosing a. Hence, h_1 and h_2 are information sets for player 2, whereas h_3 is an information set for player 1.

By using the backwards dominance procedure, it may be verified that player 2 can choose strategies (c, g) and (c, h) under common belief in future rationality. Namely, the only strategies

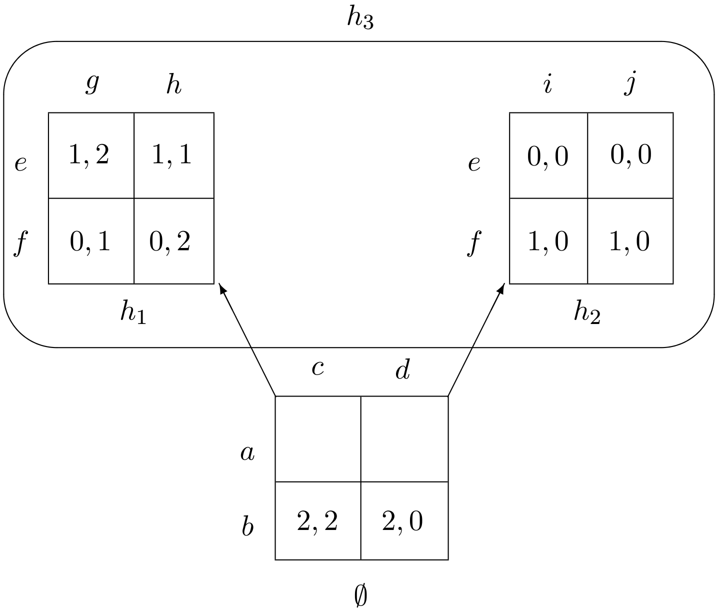

Figure 2: Common belief in future rationality does not imply sequential rationalizability

---

that can be eliminated in this procedure are strategies $(a, e), (a, f), (d, i)$ and (d, j) at $\Gamma^0(\emptyset)$, after which the procedure stops.

Under sequential rationalizability, however, player 2 can only choose strategy (c, g). Namely, at the beginning of the game, player 1 can only assign positive probability to strategies (c, g) and (c, h) since he believes in player 2's future rationality at $\emptyset$. Sequential rationalizability, however, requires that player 1's conditional beliefs are Kreps-Wilson consistent, and hence should satisfy Bayesian updating. As such, player 1 should at h_3 assign probability zero to player 2's strategies (d, i) and (d, j), and therefore player 1 should choose e at h_3. Under sequential rationalizability, player 2 should therefore believe at h_1 that player 1 chooses e, and hence player 2 should choose g at h_1, and not h. Hence, under sequential rationalizability player 2 can only rationally choose (c, g).

The reason for the difference in this example is that sequential rationalizability imposes (common belief in) Bayesian updating, whereas common belief in future rationality does not. Hence, imposing Bayesian updating would really change the concept of common belief in future rationality. This is in contrast with findings in Shimoji and Watson (1998), who have shown that for the concept of extensive form rationalizability it is inessential whether one imposes (common belief in) Bayesian updating or not – the set of strategies selected will remain the same.

In fact, from the conditions (d), (e) and (f) above it is clear that (common belief in) Kreps-Wilson consistency, together with independent beliefs about the opponents' future behavior, is the *only* difference between common belief in future rationality and sequential rationalizability.

6.2. Backwards Rationalizability

Independently from our paper, Penta (2009) has developed a procedure, *backwards rationalizability,* which is tightly related to our backwards dominance procedure. Penta's procedure restricts at every round the possible strategies and conditional belief vectors for the players, and can be described as follows.

Algorithm 6.2. *(Backwards rationalizability)*

Initial step. *For every player i, let B_i^0 be the set of all conditional belief vectors* ***satisfying Bayesian updating****, and at every information set $h \in H$, let $S_i^0 := S_i(h)$ be the set of all strategies that possibly reach h.*

Inductive step. *Let $k \geq 1$, and suppose that B_i^{k-1} and $S_i^{k-1}(h)$ have been defined for all players i, and all $h \in H$. Then, B_i^k contains those conditional belief vectors $(b_i(h))_{h \in H_i}$ in B_i^{k-1} such that $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ for all $h \in H_i$. At every information set $h \in H$, the set $S_i^k(h)$ contains those strategies $s_i \in S_i^{k-1}(h)$ that are optimal, for some conditional belief vector in B_i^k, at every $h' \in H_i(s_i)$ weakly following h.*

Let $S_i^\infty(h) := \cap_k S_i^k(h)$ be the set of strategies for player i that survive the procedure at information set h. A strategy s_i is called *backwards rationalizable* if $s_i \in S_i^\infty(\emptyset)$.

By construction, the sets $(S_i^\infty(h))_{i \in I, h \in H}$ have the following property: A strategy $s_i \in S_i(h)$ is in $S_i^\infty(h)$ if and only if there is a conditional belief vector $(b_i(h'))_{h' \in H_i}$ such that (a) $b_i(h') \in \Delta(S_{-i}^\infty(h'))$ for all $h' \in H_i$, (b) $(b_i(h'))_{h' \in H_i}$ satisfies Bayesian updating, and (c) at every $h' \in H_i(s_i)$ weakly following h, strategy s_i is optimal for $b_i(h')$.

In particular it follows that the collection $(S_i^\infty(h))_{i \in I, h \in H}$ of strategy subsets is closed under belief in future rationality. Hence, by our Theorem 5.3 we may conclude that every strategy which is backwards rationalizable can also rationally be chosen under common belief in future rationality. In fact, the only difference between the two concepts is that backwards rationalizability requires (common belief in) Bayesian updating, whereas common belief in future rationality does not. Namely, if we would drop the Bayesian updating condition (b) above, then we would obtain precisely the definition of a collection of strategy subsets that is closed under belief in future rationality. So, backwards rationalizability is weaker than sequential rationalizability, but stronger than common belief in future rationality, in terms of strategies being selected.

6.3. 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 Pearce's procedure. Both procedures refine at every round the sets of strategies and conditional beliefs of the players, and work as follows.

We start with the set of *all* strategies and conditional beliefs for each player. At every further round k we look at those information sets that can be reached by strategy profiles that have survived the previous round $k-1$. At every such information set, we restrict to conditional beliefs that assign positive probability only to opponents' strategies that have survived round $k-1$. If an information set cannot be reached by strategy profiles that have survived so far, then we impose no further restrictions on the conditional beliefs there. At round k, we then restrict to strategies that are optimal, at every information set, for conditional beliefs that have survived this round k. And so on. The strategies that survive at the end are called *extensive form rationalizable.*

Call a strategy *rational* if it is optimal, at every information set, for some conditional belief. The main idea in extensive form rationalizability can then be expressed as follows: At every information set the corresponding player first asks whether this information set can be reached by rational strategies. If so, then at that information set he must only assign positive probability to rational opponents' strategies. In that case, he then asks: Can this information set also be reached by opponents' strategies that are rational if the opponents believe, whenever possible, that their opponents choose rationally? If so, then at that information set he must only assign positive probability to such opponents' strategies. And so on. So, in a sense, at every information set the associated player looks for the highest degree of mutual belief in rationality that makes reaching this information set possible, and his beliefs at that information set should reflect this highest degree. Battigalli and Siniscalchi (2002) have formalized this argument within an epistemic model, and show that it leads precisely to the set of extensive form rationalizable strategies in every game.

In this section we wish to compare our notion of common belief in future rationality to the concept of extensive form rationalizability. To do so 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 backwards dominance algorithm, and therefore easier to compare.

Shimoji and Watson's procedure is similar in spirit to our backwards dominance procedure, as it iteratedly removes strategies from conditional games. However, their criterion for removing a strategy in a particular conditional game is different. Remember that in our backwards dominance procedure we remove a strategy for player i in the conditional game at h whenever it is strictly dominated in some conditional game for player i that *weakly follows* h. In Shimoji and Watson's procedure we remove a strategy for player i at the conditional game at h if there is some conditional game for player i, *not necessarily weakly following* h, at which it is strictly dominated. So, in Shimoji and Watson's procedure we would remove strategy s_i at h also if it is strictly dominated at some conditional game for player i which comes *before* h. Formally, their procedure can be formulated as follows.

Algorithm 6.3. *(Shimoji and Watson's iterated conditional dominance procedure)*

Initial step. *For every information set h, let $\Gamma^0(h)$ be the full conditional game at h.*

Inductive step. *Let $k \geq 1$, and suppose that the conditional games $\Gamma^{k-1}(h)$ have been defined for every information set h. Then, at every information set h delete from the conditional game $\Gamma^{k-1}(h)$ those strategies for player i that are strictly dominated within some conditional game $\Gamma^{k-1}(h')$ for player i,* ***not necessarily weakly following*** *h. This yields the new conditional games $\Gamma^k(h)$. Continue this procedure until no further strategies can be eliminated in this way.*

A strategy s_i is said to survive this procedure if $s_i \in \Gamma^k(\emptyset)$ for all k. Shimoji and Watson (1998) have shown that this procedure delivers exactly the set of extensive form rationalizable strategies. Note that in the iterated conditional dominance procedure, it is possible that at a given conditional game $\Gamma^{k-1}(h)$ *all* strategies of a player i will be eliminated in step k – something that can never happen in the backwards dominance procedure. Consider, namely, some information set $h \in H_i$, and some information set h' following h. Then, it is possible that within the conditional game $\Gamma^{k-1}(h)$, all strategies for player i in $\Gamma^{k-1}(h')$ are strictly dominated. In that case, we would eliminate in $\Gamma^{k-1}(h')$ all remaining strategies for player i! Whenever this occurs, it is understood that at every further step nothing can be eliminated from the conditional game at h' anymore.

To illustrate this important aspect, let us consider the game from Figure 1, 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 strategies for player 1 left at h_1, and hence we cannot eliminate any more strategies for player 2 at h_1 after step 1.

Now, what can we say about the relationship between common belief in future rationality and extensive form rationalizability? To answer this question, we compare the outputs of the backwards dominance procedure and the iterated conditional dominance procedure. It turns out that in terms of *strategies,* there is no logical relationship between the two concepts. Consider, to that purpose, the game in Figure 3. The full conditional games at $\emptyset$ and h_1 are represented in Table 4.

The backwards 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 conditional game $\Gamma^0(\emptyset)$, 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 conditional game $\Gamma^0(h_1)$. 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, only strategies b and e can rationally be chosen under common belief in future rationality.

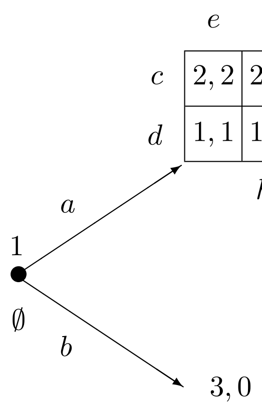

Figure 3: There is no logical relationship, in terms of strategies, between common belief in future rationality and extensive form rationalizability

---

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)	$1,1$	$1,2$	$4,0$

Table 4: The full conditional games in Figure 4

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 conditional game $\Gamma^0(\emptyset)$, 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 conditional game $\Gamma^0(h_1)$. 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 $\Gamma^1(\emptyset)$, 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 he 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.

Note that the first argument is basically a backwards induction argument, and that the second argument is a forward induction argument, leading to opposite choices for player 2. In fact, the backwards induction and forward induction flavor of both concepts is nicely illustrated by their associated algorithms: In the backwards dominance algorithm, whenever we find a strategy that is strictly dominated at some information set h, we will eliminate it at all *previous* information sets as well. This procedure thus works *backwards.* In the iterated conditional dominance procedure, on the other hand, we would then eliminate this strategy at *all* other information sets, so also at *future* information sets. This procedure thus works *backwards* and *forward.*

The game in Figure 3 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 that I do not know. My conjecture is that in every game, the set of outcomes reachable under extensive form rationalizability is always a subset of the set of outcomes reachable under common belief in future rationality. However, I have no formal proof for this conjecture. So, I leave this question as an interesting open problem here.

In Table 5 we summarize the backwards dominance procedure and the iterated conditional dominance procedure, clearly showing the differences and similarities between the two algorithms.

	At inf. set h, eliminate a strategy s_i from $\Gamma^{k-1}(h)$ if ...
Backwards dominance	s_i is strictly dominated at some conditional game $\Gamma^{k-1}(h')$ for player i that weakly follows h.
Iterated conditional dominance	s_i is strictly dominated at some conditional game $\Gamma^{k-1}(h')$ for player i, not necessarily weakly following h.

Table 5: Comparison between the two procedures

7. 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? We leave these questions for future research.

8. 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.

8.1. Some Preparatory Results

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

Definition 8.1. *(Forward inclusion property) The collection $(D_{-i}(h))_{h \in H_i}$ of strategy subsets $D_{-i}(h) \subseteq S_{-i}(h)$ satisfies the forward inclusion property if for every $h, h' \in H_i$ where h' follows h, it holds that $D_{-i}(h) \cap S_{-i}(h') \subseteq D_{-i}(h')$.*

Lemma 8.2. *(Existence of sequentially optimal strategies) For a given player i, consider a collection $(D_{-i}(h))_{h \in H_i}$ of strategy subsets satisfying the forward inclusion property. At a given information set $h^* \in H_i$ fix some conditional belief $b_i(h^*) \in \Delta(D_{-i}(h^*))$. Then, $b_i(h^*)$ can be extended to a conditional belief vector $(b_i(h))_{h \in H_i}$ on $(D_{-i}(h))_{h \in H_i}$, such that there is some strategy $s_i \in S_i(h^*)$ which is optimal at every $h \in H_i$ weakly following h^* for the belief $b_i(h)$.*

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

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

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

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

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

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

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

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

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

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

$$u_i((c_i(h), (c_i(h'))_{h' \in H_i(h)}), b_i(h)) \geq u_i((c_i', (c_i(h'))_{h' \in H_i(h)}), b_i(h)) \tag{8.1}$$

for all $c_i' \in C_i(h)$. Here, $H_i(h)$ denotes the collection of information sets in H_i that follow h. In this way, we specify at every $h \in H_i$ weakly following h^* a choice $c_i(h)$ that satisfies (8.1).

Now, let s_i be the strategy that
(a) at every $h \in H_i(s_i)$ weakly following h^*, prescribes the optimal choice $c_i(h)$ as in (8.1),
(b) at every $h \in H_i(s_i)$ preceding h^*, prescribes the unique choice $c_i(h)$ that leads to h^*, and
(c) at every other $h \in H_i(s_i)$ specifies an arbitrary choice.

By construction the strategy s_i is in $S_i(h^*)$, as it prescribes all choices that lead to h^*. As the conditional belief vector $(b_i(h))_{h \in H_i}$ satisfies Bayesian updating at information sets weakly following h^*, 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^*. That is, every strategy s_i for which the choices $c_i(h)$ are optimal in the sense of (8.1), is optimal as a strategy at every $h \in H_i$ weakly following h^*. Hence, we may conclude that the strategy s_i so constructed is optimal at every $h \in H_i(s_i)$ weakly following h^* for the belief $b_i(h)$. Since s_i is in $S_i(h^*)$, and $(b_i(h))_{h \in H_i}$ is a conditional belief vector on $(D_{-i}(h))_{h \in H_i}$ which extends $b_i(h^*)$, the proof is complete. ■

The lemma above implies in particular that, whenever the collection $(D_{-i}(h))_{h \in H_i}$ satisfies the forward inclusion property, then it allows for a conditional belief vector $(b_i(h))_{h \in H_i}$ and a strategy s_i, such that s_i is optimal at every $h \in H_i(s_i)$ for the belief $b_i(h)$. In other words, collections $(D_{-i}(h))_{h \in H_i}$ 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 backwards dominance procedure satisfy the forward inclusion property. This result thus guarantees that we can apply Lemma 8.2 to every round of the backwards dominance procedure – something that will be important for proving some of our theorems in the paper.

Lemma 8.3. *(Backwards dominance procedure satisfies forward inclusion property) For every information set h and player i, let $S_i^k(h)$ be the set of player i strategies in $\Gamma^k(h)$ – the conditional* *game at h produced in round k of the backwards dominance procedure. Then, the collection $(S_{-i}^k(h))_{h\in H_i}$ of strategy subsets satisfies the forward inclusion property.*

Proof. For $k=0$ the statement is trivial since $S_{-i}^0(h)=S_{-i}(h)$ for all h. So, take some $k\geq 1$. Suppose that $h,h'\in H_i$ and that h' follows h. Take some opponent's strategy s_j in $S_{-i}^k(h)\cap S_{-i}(h')$, that is, $s_j\in S_j^k(h)\cap S_j(h')$. Then, since $s_j\in S_j^k(h)$, we have that s_j is not strictly dominated in any conditional game $\Gamma^{k-1}(h'')$ where $h''\in H_j(s_j)$ weakly follows h. As h' follows h, it holds in particular that s_j is not strictly dominated in any conditional game $\Gamma^{k-1}(h'')$ where $h''\in H_j(s_j)$ weakly follows h'. Together with the fact that $s_j\in S_j(h')$, this implies that $s_j\in S_j^k(h')$. So, $S_{-i}^k(h)\cap S_{-i}(h')\subseteq S_{-i}^k(h')$, and hence the forward inclusion property holds. ■

Our third lemma shows an important optimality property of our backwards dominance procedure. Recall that in the backwards dominance procedure, $\Gamma^k(h)$ denotes the conditional game at h produced at the end of round k. For every player i, let us denote by $S_i^k(h)$ the set of player i strategies in $\Gamma^k(h)$. By construction of the algorithm, $S_i^k(h)$ contains exactly those strategies in $S_i^{k-1}(h)$ that, at every $h'\in H_i(s_i)$ weakly following h, are not strictly dominated in $\Gamma^{k-1}(h')$. By Lemma 3 in Pearce (1984), we know that s_i is not strictly dominated in $\Gamma^{k-1}(h')$ if and only if there is some belief $b_i(h')\in\Delta(S_{-i}^{k-1}(h'))$ such that s_i is optimal for $b_i(h')$ *among all strategies in* $S_i^{k-1}(h')$. That is,

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

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

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

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

Lemma 8.4. *(Optimality principle for backwards dominance procedure) Let $S_i^k(h)$ denote the set of player i strategies in the conditional game $\Gamma^k(h)$ produced in round k of the backwards dominance procedure. Then, $s_i\in S_i^k(h)$ if and only if for every $h'\in H_i(s_i)$ weakly following h there is some belief $b_i(h')\in\Delta(S_{-i}^{k-1}(h'))$ such that s_i is optimal for $b_i(h')$ among all strategies in $S_i(h')$.*

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 $h'\in H_i(s_i)$ weakly following h. Then we know from our argument above that there is some $b_i(h')\in\Delta(S_{-i}^{k-1}(h'))$ such that

$$u_i(s_i,b_i(h'))\geq u_i(s_i',b_i(h')) \text{ for all } s_i'\in S_i^{k-1}(h'). \tag{8.2}$$

We will prove that, in fact,

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

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

$$u_i(s_i, b_i(h')) < u_i(s_i', b_i(h')). \tag{8.3}$$

We show that in this case there would be some $s_i^* \in S_i^{k-1}(h')$ with $u_i(s_i', b_i(h')) \leq u_i(s_i^*, b_i(h'))$, which together with (8.3) would contradict (8.2).

From Lemma 8.3 we know that the collection $(S_{-i}^{k-1}(h''))_{h'' \in H_i}$ satisfies the forward inclusion property. Hence, by Lemma 8.2, we can extend $b_i(h')$ to some conditional belief vector $(b_i(h''))_{h'' \in H_i}$ with $b_i(h'') \in \Delta(S_{-i}^{k-1}(h''))$ for all $h'' \in H_i$, and we can find some strategy $s_i^* \in S_i(h')$ which is optimal, at every $h'' \in H_i(s_i^*)$ weakly following h', for the belief $b_i(h'')$. But then, it follows that $s_i^* \in S_i^k(h')$, and hence in particular $s_i^* \in S_i^{k-1}(h')$. Moreover, s_i^* is optimal at h' for the belief $b_i(h')$. And hence, we have by (8.3) that

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

This, however, contradicts (8.2). So, (8.3) must be incorrect, and hence s_i is optimal at h' for the belief $b_i(h')$ among all strategies in $S_i(h')$. ■

8.2. Backwards Dominance Procedure Delivers Nonempty Output

We now prove Theorem 4.2, which states that the backwards dominance procedure delivers at every information set a conditional game with nonempty strategy sets. Recall that $S_i^k(h)$ denotes the set of player i strategies in the conditional game $\Gamma^k(h)$ produced by round k of the backwards 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 $k \geq 1$, 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 8.3 we know that the collection $(S_{-i}^{k-1}(h))_{h \in H_i}$ satisfies the forward inclusion property. Hence, by Lemma 8.2, we can find a conditional belief vector $(b_i(h))_{h \in H_i}$ with $b_i(h) \in \Delta(S_{-i}^{k-1}(h))$ for all $h \in H_i$, and a strategy $s_i \in S_i(h^*)$, such that s_i is optimal at every $h \in H_i$ weakly following h^* for the belief $b_i(h)$. But then, we know from Lemma 8.4 that $s_i \in S_i^k(h^*)$, and hence $S_i^k(h^*)$ is nonempty. By induction on k, the proof is complete. ■

8.3. Backwards Dominance Procedure Works

We now prove our main result, Theorem 4.3, which states that the backwards dominance procedure yields exactly those strategies that can rationally be chosen under common belief in future rationality. We thus must prove two directions: First, that every strategy that can rationally be chosen under common belief in future rationality survives the backwards 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 survives the backwards dominance procedure.

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

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

So, $B_i(h)$ contains those conditional beliefs at h about the opponents' strategy choices that are possible under common belief in future rationality. Recall that $S^k_{-i}(h)$ denotes the set of opponents' strategies in the conditional game $\Gamma^k(h)$ produced in round k of the backwards dominance procedure. We prove the following claim.

Claim. $B_i(h) \subseteq \Delta(S^k_{-i}(h))$ for every k.

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

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

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

As t_i expresses common belief in future rationality, $b_i(t_i, h)$ only assigns positive probability to opponents' strategy-type pairs (s_j, t_j) where s_j is optimal for t_j at every $h' \in H_j(s_j)$ weakly following h. Together with the fact that $b_i(t_i, h)$ only assigns positive probability to opponents' types t_j whose belief at such h' about the other players' strategy choices is in $\Delta(S^{k-1}_{-j}(h'))$, this implies that $b_i(t_i, h)$ only assigns positive probability to opponents' strategies s_j that are optimal, at every $h' \in H_j(s_j)$ weakly following h, for some belief in $\Delta(S^{k-1}_{-j}(h'))$. However, by Lemma 8.4, these latter strategies s_j are exactly the strategies in $S^k_j(h)$. Hence, $b_i(t_i, h)$ only assigns positive probability to opponents' strategies in $S^k_j(h)$, which means that the marginal of $b_i(t_i, h)$ on $S_{-i}(h)$ is in $\Delta(S^k_{-i}(h))$. By definition, the marginal of $b_i(t_i, h)$ on $S_{-i}(h)$ was $b_i(h)$, so $b_i(h) \in \Delta(S^k_{-i}(h))$.

Since this holds for every $b_i(h) \in B_i(h)$, we may conclude that $B_i(h) \subseteq \Delta(S^k_{-i}(h))$. 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 $h \in H_i(s_i)$ for the belief $b_i(t_i, h)$. By the claim above we know that $b_i(t_i, h) \in \Delta(S_{-i}^{\infty}(h))$, where $S_{-i}^{\infty}(h) := \cap_k S_{-i}^k(h)$. So, at every $h \in H_i(s_i)$ strategy s_i is optimal for some belief in $\Delta(S_{-i}^{\infty}(h))$. By Lemma 8.4 this implies that $s_i \in S_i^{\infty}(\emptyset)$, where $S_i^{\infty}(\emptyset) := \cap_k S_i^k(\emptyset)$. This means, however, that s_i survives the backwards dominance procedure, and hence the proof of part (a) is complete.

(b) Every strategy that survives the backwards dominance procedure can rationally be chosen under common belief in future rationality.

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 backwards dominance procedure. So, $S_i^{\infty}(h) := \cap_k S_i^k(h)$. Remember that $S_i^{\infty}(\emptyset)$ contains exactly those player i strategies that survive the backwards 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) set of histories $h \in H_i$ for which $s_i \in S_i^{\infty}(h)$. So, by Lemma 8.4, we can find for every $s_i \in S_i$ some conditional belief vector $(b_i(s_i, h))_{h \in H_i}$ such that (a) $b_i(s_i, h) \in \Delta(S_{-i}^{\infty}(h))$ for every $h \in H_i$, and (b) s_i is optimal at every $h \in H_i^*(s_i)$ for the belief $b_i(s_i, h)$.

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

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

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

$b_i(t_i^{s_i}, h)$ only assigns positive probability to opponents' strategy-type pairs $(s_j, t_j^{s_j})$ where $s_j \in S_j^\infty(h)$.

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

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

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

8.4. Best-Response Characterization

We finally prove Theorem 5.3, 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. Recall that $S_i^\infty(h)$ denotes the set of player i strategies that are part of the conditional game at h at the end of the backwards dominance procedure. Then, from Lemma 8.4 it immediately follows that the collection of strategy subsets $(S_i^\infty(h))_{h \in H, i \in I}$ is closed under belief in future rationality. Since s_i can rationally be chosen under common belief in future rationality, we know from our Theorem 4.3 that s_i survives the backwards dominance procedure, so $s_i \in S_i^\infty(\emptyset)$. Hence, the collection $(S_i^\infty(h))_{h \in H, i \in I}$ of strategy subsets is closed under belief in future rationality and $s_i \in S_i^\infty(\emptyset)$, which completes the proof of the first direction.

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 in the conditional game $\Gamma^k(h)$ produced in round k of the backwards dominance procedure.

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

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

Take now some $k \geq 1$, 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 for every $h' \in H_i(s_i)$ weakly following h be some belief $b_i(h') \in \Delta(D_{-i}(h'))$ for which s_i is optimal. As, by induction assumption, $D_{-i}(h') \subseteq S_{-i}^{k-1}(h')$, there is for every $h' \in H_i(s_i)$ weakly following h some belief $b_i(h') \in \Delta(S_{-i}^{k-1}(h'))$ for which s_i is optimal. But then, by our Lemma 8.4, $s_i \in S_i^k(h)$. 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 dominance 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 5.3. ■

References

[1] Asheim, G.B. (2002), On the epistemic foundation for backward induction, *Mathematical Social Sciences* **44,** 121-144.

[2] Asheim, G.B. and A. Perea (2005), Sequential and quasi-perfect rationalizability in extensive games, *Games and Economic Behavior* **53,** 15-42.

[3] Battigalli, P. (1997), On rationalizability in extensive games, *Journal of Economic Theory* **74,** 40-61.

[4] Battigalli, P. and M. Siniscalchi (2002), Strong belief and forward induction reasoning, *Journal of Economic Theory* **106,** 356-391.

[5] Bernheim, B.D. (1984), Rationalizable strategic behavior, *Econometrica* **52,** 1007-1028.

[6] Dekel, E., Fudenberg, D. and D.K. Levine (1999), Payoff information and self-confirming equilibrium, *Journal of Economic Theory* **89**, 165–185.

[7] Dekel, E., Fudenberg, D., and D.K. Levine (2002), Subjective uncertainty over behavior strategies: A correction, *Journal of Economic Theory* **104,** 473–478.

[8] Feinberg, Y. (2005), Subjective reasoning–dynamic games, *Games and Economic Behavior* **52,** 54-93.

[9] Kreps, D.M. and R. Wilson (1982), Sequential equilibria, *Econometrica* **50,** 863–94.

[10] Pearce, D.G. (1984), Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52,** 1029-1050.

[11] Penta, A. (2009), Robust dynamic mechanism design, Manuscript, University of Pennsylvania.

[12] Perea A. (2002), A note on the one-deviation property in extensive games, *Games and Economic Behavior* **40,** 322-338.

[13] Perea, A. (2006), Proper belief revision and rationalizability in dynamic games, *International Journal of Game Theory* **34,** 529-559.

[14] Perea, A. (2007a), Proper belief revision and equilibrium in dynamic games, *Journal of Economic Theory* **136,** 572-586.

[15] Perea, A. (2007b), Epistemic foundations for backward induction: An overview, *Interactive Logic Proceedings of the 7th Augustus de Morgan Workshop, London. Texts in Logic and Games 1* (Johan van Benthem, Dov Gabbay, Benedikt Löwe (eds.)), Amsterdam University Press, 159-193.

[16] Samet, D. (1996), Hypothetical knowledge and games with perfect information, *Games and Economic Behavior* **17,** 230-251.

[17] Shimoji, M. and J. Watson (1998), Conditional dominance, rationalizability, and game forms, *Journal of Economic Theory* **83,** 161-195.

[18] Tan, T. and S.R.C. Werlang (1988), The Bayesian foundations of solution concepts of games, *Journal of Economic Theory* **45,** 370-391.