Forward Induction Reasoning versus Equilibrium Reasoning[*]

EPICENTER Working Paper No. 5 (2015)

Andrés Perea
Maastricht University

This version: April 2015

Abstract

In the literature on static and dynamic games, most rationalizability concepts have an equilibrium counterpart. In two-player games, the equilibrium counterpart is obtained by taking the associated rationalizability concept and adding the following *correct beliefs assumption:* (a) a player believes that the opponent is correct about his beliefs, and (b) a player believes that the opponent believes that he is correct about the opponent's beliefs. This paper shows that there is no equilibrium counterpart to the forward induction concept of *extensive-form rationalizability* (Pearce (1984), Battigalli (1997)), epistemically characterized by *common strong belief in rationality* (Battigalli and Siniscalchi (2002)). The reason is that there are games where the epistemic conditions of common strong belief in rationality are logically inconsistent with the correct beliefs assumption. In fact, we show that this inconsistency holds for "most" dynamic games of interest.

JEL Classification: C72

[*] I thank Christian Bach for some very valuable comments on this paper.

1 Introduction

Roughly speaking, the concepts that are used nowadays to analyze games can be divided into two categories: *equilibrium concepts* and *rationalizability concepts.* Historically, the equilibrium concepts came first, starting with the concept of Nash equilibrium (Nash (1950, 1951)), and followed by refinements such as perfect equilibrium (Selten (1975)) and proper equilibrium (Myerson (1978)) for static games, and subgame perfect equilibrium (Selten (1965)) and sequential equilibrium (Kreps and Wilson (1982)) for dynamic games, among others. It was only in the early eighties when rationalizability concepts systematically entered the game-theoretic picture, triggered by the pioneering work of Bernheim (1984), Pearce (1984) and Brandenburger and Dekel (1987), who developed the concept of *rationalizability.* Later, various refinements of rationalizability have been proposed and studied, both for static and dynamic games. Table 1 below gives an overview of these refinements.

But what precisely is it that distinguishes rationalizability concepts from equilibrium concepts? *Epistemic game theory* has given a formal answer to this question. Essentially, the difference is that equilibrium concepts make a "correct beliefs assumption", stating that a player believes that his opponents are correct about the beliefs he holds, and that a player believes that every opponent i believes that the other players are correct about i's beliefs as well. Rationalizability concepts, on the other hand, do not make this correct beliefs assumption.

As an illustration, let us compare the concepts of rationalizability and Nash equilibrium. For two-player games, rationalizability can be characterized epistemically by *common belief in rationality* (Brandenburger and Dekel (1987), Tan and Werlang (1988)), stating that both players believe that both players choose rationally, both players believe that both players believe that both players choose rationally, and so on. On the other hand, Brandenburger and Dekel (1987, 1989), Tan and Werlang (1988), Aumann and Brandenburger (1995), Asheim (2006) and Perea (2007) have shown that Nash equilibrium in two-player games can be characterized epistemically by common belief in rationality[1], together with the correct beliefs assumption above. Hence, for two-player games it is the correct beliefs assumption – and nothing more – that separates rationalizability from Nash equilibrium. For that reason, we may view Nash equilibrium as the *equilibrium counterpart* to rationalizability. See Table 1. For games with more than two players, additional conditions beyond the correct beliefs assumption are needed to close the gap between rationalizability and Nash equilibrium. We refer to the above mentioned papers, together with Barelli (2009) and Bach and Tsakas (2014), for the details.

In fact, a similar relationship holds between various other rationalizability and equilibrium concepts. For instance, the equilibrium concept of perfect equilibrium is obtained in two-player games if we assume the epistemic conditions underlying permissibility (Brandenburger (1992),

[1] In fact, for the epistemic characterization of Nash equilibrium not all the conditions in *common belief in rationality* are needed. If we make the correct beliefs assumption, then *mutual belief in rationality* – stating that both players believe that both players choose rationally – already implies common belief in rationality, and hence is sufficient to characterize Nash equilibrium. See Polak (1999).

Rationalizability concept	**Equilibrium counterpart**
Static games	
Rationalizability (Bernheim (1984), Pearce (1984), Brandenburger and Dekel (1987)) **Common belief in rationality** (Tan and Werlang (1988))	**Nash equilibrium** (Nash (1950, 1951)))
Permissibility (Brandenburger (1992), Börgers (1994))	**Perfect equilibrium** (Selten (1975))
Proper rationalizability (Schuhmacher (1999), Asheim (2001))	**Proper equilibrium** (Myerson (1978))
Iterated elimination of weakly dominated choices **Iterated assumption of rationality** (Brandenburger, Friedenberg and Keisler (2008))	**?**
Dynamic games	
Common belief in future rationality (Perea (2014), Baltag, Smets and Zvesper (2009), Penta (2009)) **Sequential rationalizability** (Dekel, Fudenberg and Levine (1999, 2002), Asheim and Perea (2005))	**Subgame perfect equilibrium** (Selten (1965) **Sequential equilibrium** (Kreps and Wilson (1982))
Extensive-form rationalizability (Pearce (1984), Battigalli (1997)) **Common strong belief in rationality** (Battigalli and Siniscalchi (2002))	**None** See this paper

Table 1: Rationalizability concepts and their equilibrium counterparts

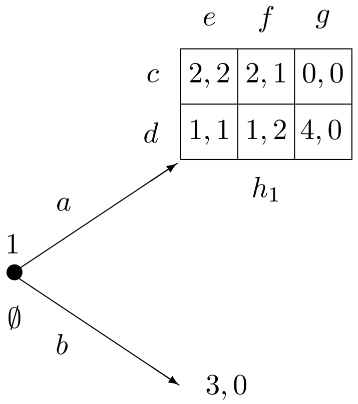

Figure 1: Extensive-form rationalizability is inconsistent with equilibrium reasoning

Börgers (1994)), and on top of this impose the correct beliefs assumption. See Perea (2012, Problem 5.9). As such, we may view perfect equilibrium as the equilibrium counterpart to the rationalizability concept of permissibility. In Table 1 we give an overview of several other rationalizability concepts together with their equilibrium counterparts.

One thing we learn from Table 1 is that there is an equilibrium counterpart to *most* of the rationalizability concepts that have been studied in the literature. Indeed, if we choose any rationalizability concept from rationalizability, permissibility, proper rationalizability (Schuhmacher (1999), Asheim (2001)), common belief in future rationality (Perea (2014), Baltag, Smets and Zvesper (2009), Penta (2009)) or sequential rationalizability (Dekel, Fudenberg and Levine (1999, 2002), Asheim and Perea (2005)), and combine the underlying epistemic conditions with the correct beliefs assumption, then we obtain a well-defined equilibrium concept in the right-hand column of Table 1 (at least for two-player games).

But is this true for *all* rationalizability concepts in Table 1? The answer is "no": For the forward induction concept of *extensive-form rationalizability* (Pearce (1984), Battigalli (1997)), epistemically characterized by *common strong belief in rationality* (Battigalli and Siniscalchi (2002)), we show that there is *no* equilibrium counterpart to it.[2] This is actually the main message of this paper. The reason, as we will see, is that the epistemic conditions of *common strong belief in rationality* in a belief-complete type structure are *logically inconsistent* with the correct beliefs assumption in some – in fact, most – dynamic games.

To obtain an informal understanding of this result, consider the dynamic game in Figure 1. At the beginning of the game, $\emptyset$, player 1 can choose between a and b. If he chooses b, the game

[2] For iterated elimination of weakly dominated choices, epistemically characterized by iterated assumption of rationality, we do not know whether an equilibrium conterpart exists. Our intuition suggests that it does not exist, but this remains to be investigated. This, however, is beyond the scope of this paper.

ends and the utilities for players 1 and 2 would be 3 and 0, respectively. If he chooses a then the game moves to history h_1, where player 1 can choose between c and d, and player 2 can simultaneously choose between e, f and g.

The main condition in extensive-form rationalizability – and this is in fact the first layer in common strong belief in rationality – is that a player, at each of his information sets, must believe that his opponent is choosing rationally, whenever it is possible to believe so. We say that players must *strongly believe in the opponent's rationality.* This is a typical *forward induction* argument, and hence the concept is generally viewed as a typical forward induction concept. Moreover, every player i must believe – whenever this is possible – that opponent j does not only choose rationally, but also that j strongly believes in i's rationality. This is the second layer in common strong belief in rationality. We will now argue that these two conditions together are *incompatible* with the correct beliefs assumption in the game above.

If player 1 strongly believes in 2's rationality, then player 1 will certainly believe that player 2 will *not* choose g, as e and f are always better than g for player 2 at history h_1. At the same time, if player 2 strongly believes in 1's rationality, then player 2 will believe at h_1 that choosing a was actually a conscious, optimal choice for player 1. However, this would require player 2 to believe, at h_1, that player 1 assigns a high probability to player 2's irrational choice g, since otherwise a can never be an optimal choice for player 1.

Consequently, if player 1 believes that player 2 strongly believes in 1's rationality, as we assume, then player 1 must believe that player 2 will believe, at h_1, that player 1 assigns a high probability to player 2's choice g.

The first two layers in common strong belief in rationality thus imply that player 1, on the one hand, assigns probability 0 to player 2's irrational choice g, but that player 1, at the same time, believes that player 2 will believe at h_1 that player 1 assigns a high probability to player 2 choosing g. In other words, player 1 must believe that player 2, at h_1, will be *wrong* about player 1's actual beliefs. But this means that the correct beliefs assumption cannot hold. That is, common strong belief in rationality – and hence extensive-form rationalizability – is logically inconsistent with the correct beliefs assumption in the game of Figure 1. In Section 6 we will actually show that the game in Figure 1 is not a coincidence: in "most" dynamic games, extensive-form rationalizability will be inconsistent with the correct beliefs assumption.

Despite this inconsistency between the forward induction reasoning in common strong belief in rationality on the one hand, and equilibrium reasoning on the other hand, there are forward induction *equilibrium* concepts in the literature where equilibrium reasoning is *imposed* on the players. Examples are *justifiable sequential equilibrium* (McLennan (1985)), *Cho's forward induction equilibrium* (Cho (1987)), *stable sets of beliefs* (Hillas (1994)), *explicable equilibrium* (Reny (1992)), *outcomes satisfying forward induction* (Govindan and Wilson (2009)) and *Man's forward induction equilibrium* (Man (2012)). These concepts, in contrast to common strong belief in rationality, impose the correct beliefs assumption as an *exogenous restriction* on the players' belief hierarchies. This means that players are not only assumed to hold belief hierarchies that satisfy the correct beliefs assumption, but are also restricted to attribute "unexpected" moves by the opponent to opponent's belief hierarchies that satisfy the correct beliefs assumption. That is, players are restricted to reason entirely within the boundaries set by the correct beliefs assumption. As an illustration, take the concept of justifiable sequential equilibrium (McLennan (1985)), which is defined as a refinement of sequential equilibrium. Within this concept, players are not only assumed to hold belief hierarchies that correspond to a sequential equilibrium – and hence, in particular, satisfy the correct beliefs assumption – but in addition, when players are trying to explain an opponent's move they did not expect, they can only attribute such moves to opponent's belief hierarchies that also correspond to a sequential equilibrium. In other words, the reasoning of players is assumed to take place entirely within the context of sequential equilibrium. Unexpected moves cannot be explained by belief hierarchies that fall outside the boundaries set by sequential equilibrium. Similar exogenous restrictions are imposed by the other forward induction equilibrium concepts mentioned above. We refer the reader to Section 7 for the details.

Imposing the correct beliefs assumption as an exogenous restriction on the players' reasoning comes at a cost, however. We show in Section 7 that none of the forward induction equilibrium concepts above is able to uniquely select the intuitive forward induction strategy f for player 2 in the game of Figure 1. The reason is that, if player 2 at h_1 wishes to rationalize the "surprising" move a by player 1, then player 2 must believe that player 1's belief hierarchy *violates* the correct beliefs assumption – something that is "not allowed" by the forward induction equilibrium concepts above. Common strong belief in rationality, in contrast, does not impose such exogenous restrictions, and *is* therefore able to uniquely select the intuitive forward induction strategy f for player 2.

Such exogenous restrictions on the the players' belief hierarchies in forward induction reasoning have been explicitly studied in Battigalli and Friedenberg (2012). They take the forward induction concept of common strong belief in rationality, but do so relative to a type structure that *does not necessarily contain all belief hierarchies.* By excluding some belief hierarchies from the type structure, they thus impose some *exogenous restriction* on the players' belief hierarchies, as players can only hold – and reason about – belief hierarchies that are within the type structure. It would be interesting to see whether some of the forward induction equilibrium concepts mentioned above, which do impose exogenous restrictions on the players' belief hierarchies, can be characterized within the Battigalli-Friedenberg framework by common strong belief in rationality relative to a suitably restricted type structure. We leave this question for future research.

The outline of this paper is as follows. In Section 2 we give a formal model of dynamic games. In Section 3 we show how infinite hierarchies of conditional beliefs in dynamic games can be encoded by means of epistemic models with types. Section 4 formally introduces the notion of a *reasoning context,* which assigns to every dynamic game a collection of belief hierarchies for each of the players, and describes what it means for a reasoning context to be consistent with equilibrium reasoning. In Section 5 we define the concept of common strong belief in rationality, and show that in some dynamic games it is inconsistent with equilibrium reasoning. In Section

6 we provide a characterization of the class of two-player dynamic games for which common strong belief in rationality *is* consistent with equilibrium reasoning, and show that this class is actually very small. This is the content of Theorem 6.2. For this characterization we use a specific version of *extensive-form best response sets* as developed by Battigalli and Friedenberg (2012). In Section 7 we discuss some forward induction *equilibrium* concepts that have been proposed in the literature, and explain why these concepts, in the game of Figure 1, fail to uniquely select the intuitive forward induction strategy f for player 2. Section 8 contains the proof of Theorem 6.2.

2 Dynamic Games

2.1 A Model of Dynamic Games

In this paper, we will restrict our attention to dynamic games with *two players* and *observable past choices.* We assume moreover that the dynamic game is *finite* – that is, the game ends after finitely many moves, and every player has finitely many choices available at every moment in time where it is his turn to move. The first two restrictions are mainly for the ease of exposition. We believe that all results can be extended to more general finite dynamic games. One should bear in mind, however, that for games with more than two players the correct beliefs assumption alone is not sufficient to characterize equilibrium reasoning. Hence, additional "equilibrium" conditions are needed to turn a rationalizability concept into its equilibrium counterpart.

Formally, a *finite dynamic game G* with *two players* and *observable past choices* consists of the following ingredients.

First, there is the set of players $I = \{1, 2\}$. The instances where one or both players must make a choice are given by a finite set H of non-terminal histories. The possible instances where the game ends are described by a finite set Z of terminal histories. By $\emptyset$ we denote the beginning of the game.

Consider a non-terminal history h at which player i must make a choice. We assume that player i observes precisely which choices have been made by his opponent in the past. That is, we assume that the dynamic game is with *observable past choices.* By $C_i(h)$ we denote the finite set of choices that are available to player i at h.

We explicitly allow for *simultaneous moves* in the dynamic game. That is, we allow for non-terminal histories h at which both players 1 and 2 make a choice. By $I(h)$ we denote the set of active players at h. That is, $I(h)$ contains those players who must make a choice at h. Every combination of choices $(c_i)_{i \in I(h)}$ at h is assumed to move the game from the non-terminal history h to some other (terminal or non-terminal) history h'. These transitions can formally be described by a *move-function m,* which assigns to every non-terminal history h, and every combination of choices $(c_i)_{i \in I(h)}$, the (terminal or non-terminal) history $m(h)$ that follows. By H_i we denote the collection of non-terminal histories where player i is active.

Players are assumed to have preferences over the possible outcomes in the game, representable by utility functions over the set of terminal histories Z. Formally, for every terminal history $z \in Z$ and player i, we denote by $u_i(z)$ the utility for player i at z, representing how desirable he deems the outcome z.

In the remainder of this paper, we adopt the following conventions: Whenever we talk about a dynamic game G we will always assume that it is a finite dynamic game with two players and observable past choices. Moreover, whenever we mention a history h we will always assume that it is a non-terminal history, unless explicitly stated otherwise. Finally, when we mention players i and j in the same sentence, we always assume that $i \neq j$.

2.2 Strategies

Intuitively, a *strategy* for a player is a complete plan which describes what he will, or would, do in every situation that could possibly arise in the game. By definition, the possible situations in the game where player i must make a choice are exactly the histories in H_i. So, a possible definition of a strategy for player i – and this is in fact the traditional definition of a strategy in game theory – would be a function that assigns an available choice to *each* history where player i is active. The problem with this definition, however, is that it may contain some redundant information, as certain future histories where player i is active can be excluded by choices of player i at earlier histories. In that case, it is no longer relevant to specify what this player would do at those excluded histories, as those histories will certainly not be reached if the player implements the strategy correctly – as we suppose him to do. Consider, for instance, the game in Figure 1. If player 1 decides to go for b at the beginning of the game, he is certain that the history h_1 will not be reached. So in that case it is redundant to specify what player 1 would do were h_1 to be reached, as h_1 is clearly avoided by the choice b. We may therefore view b as a complete plan, although b is not a strategy in the traditional sense. In fact, we will accept b as a full description of a strategy for player 1.

An argument that is often used in defense of the traditional definition of a strategy is that the choices specified at precluded histories reflect the opponents' counterfactual beliefs about his future behavior if the player decides to deviate from his plan. See Rubinstein (1991) for a discussion of this issue. But this would mean that the strategy represents both choices and beliefs – something I consider highly undesirable. In my opinion, we should always clearly separate objects of choice from beliefs, and to put them in the same object is likely to cause confusion. After all, the term strategy suggests that it reflects only the plan of choices of a player. The beliefs of the players will anyhow be modelled separately in the next section, so there is no need to mix them with the players' choices.

Having said this, we opt for a definition of a strategy that only prescribes choices at those histories *not precluded* by earlier choices. To define this formally, consider two histories h and h' where player i is active, and an available choice $c \in C_i(h)$ at h. We say that choice c *avoids* history h' if h precedes h', and choosing c at h can never lead to h'.

Definition 2.1 (Strategy) *A strategy for player i is a function $s_i : \hat{H}_i \to \cup_{h \in \hat{H}_i} C_i(h)$ where (1) $\hat{H}_i \subseteq H_i$, (2) $s_i(h) \in C_i(h)$ for all $h \in \hat{H}_i$, (3) for every $h \in \hat{H}_i$ there is no preceding $h' \in \hat{H}_i$ such that the prescribed choice $s_i(h')$ avoids h, and (4) for every $h \in H_i$, if h is not avoided by any prescribed choice $s_i(h')$ with $h' \in \hat{H}_i$, then h must be in $\hat{H}_i$.*

Conditions (3) and (4) thus guarantee that $\hat{H}_i$ contains exactly those histories in H_i not precluded by earlier choices – not more and not less. The definition of a strategy we use corresponds to what Rubinstein (1991) calls a *plan of action.*

Let us denote by S_i the set of all strategies for player i. Since the dynamic game G is finite, the set S_i will be finite as well. By $S := S_1 \times S_2$ we denote the set of all strategy combinations. For a given history $h \in H$, let $S(h)$ be the set of strategy combinations that reach h. By $S_i(h)$ we denote the set of strategies s_i for player i for which there is some opponent's strategy $s_j \in S_j$ such that $(s_i, s_j) \in S(h)$. We say that strategies in $S_i(h)$ are *consistent with* h. It is not difficult to see that $S(h) = S_1(h) \times S_2(h)$ for every $h \in H$.

3 Modelling Belief Hierarchies

We now wish to model the players' beliefs in a dynamic game. There are at least two complications that we face here.

First, when players reason about their opponents in a dynamic game, they do not only hold beliefs about what other players do (first-order beliefs), but also hold second-order beliefs about the opponents' first-order beliefs about what others do, and third-order beliefs about the opponents' second-order beliefs, and so on. So, players hold a full *infinite belief hierarchy.*

Secondly, a player in a dynamic game may have to *revise* his belief if the game moves from one history to another. That is, a player will hold at each history where he is active a new conditional belief about the opponent which is compatible with the event that this particular history has been reached. Consider some player i who observes that history $h \in H_i$ has been reached. Then he knows that his opponent must be implementing some strategy in $S_j(h)$ – the set of j's strategies that make reaching h possible – and hence player i must at h restrict his belief to opponents' strategies in $S_j(h)$. And this conditional belief may be – partially or completely – contradicted at some later history, in which case he must change his belief there.

Summarizing, we see that we need to model *conditional belief hierarchies* for a player, which specify at each history where he is active what he believes about the opponent's strategy choices, about the opponent's first-order beliefs, about the opponent's second-order beliefs, and so on. But how can we model such complicated objects? One way to do so is by using a Harsanyi-style model with types (Harsanyi (1967–1968)) and adapt it to dynamic games. To see how this works, consider a player i who at history $h \in H_i$ holds a belief about the opponent's strategies, the opponent's first-order beliefs, the opponent's second-order beliefs, and so on. In other words, this player holds at h a belief about the opponent's strategies and the opponent's conditional belief hierarchies. So, a conditional belief hierarchy for player i specifies at each history in H_i a conditional belief about the opponent's strategy choices and the opponent's conditional belief hierarchies. If we substitute the word "belief hierarchy" by the word "type" then we obtain the following definition.

Definition 3.1 (Epistemic model) *Consider a dynamic game G. An epistemic model for G is a tuple $M = (T_i, b_i)_{i \in I}$ where*

(a) T_i is a set of types for player i,

(b) b_i is a function that assigns to every type $t_i \in T_i$, and every history $h \in H_i$, a probability distribution $b_i(t_i, h) \in \Delta(S_j(h) \times T_j)$.

Recall that $S_j(h)$ represents the set of opponent's strategies that are consistent with h. For every set X, we denote by $\Delta(X)$ the set of probability distributions on X with respect to some appropriately chosen σ-algebra on X. Clearly, player i must at h only assign positive probability to opponent's strategies in $S_j(h)$, as these are the only strategies compatible with the event that h is reached. This explains the condition in (b) that $b_i(t_i, h) \in \Delta(S_j(h) \times T_j)$.

By construction, at every history $h \in H_i$ type t_i holds a conditional probabilistic belief $b_i(t_i, h)$ about j's strategies and types. In particular, type t_i holds conditional beliefs about j's strategies. As each of j's types holds conditional beliefs about i's strategies, every type t_i holds at every $h \in H_i$ also a conditional belief about j's conditional beliefs about i's 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 opponent's strategies, but also about the opponent's conditional beliefs. In fact, for a given type t_i within an epistemic model, we can *derive* the complete belief hierarchy it induces.

4 Reasoning Contexts

4.1 Definition

A *reasoning context* imposes restrictions on the way a player reasons about his opponent in a dynamic game. Remember from the previous section that we have summarized the reasoning of a player by a *conditional belief hierarchy,* which describes at each history where he is active what he believes about the opponent's strategy choices, the opponent's first-order beliefs, the opponent's second-order beliefs, and so on. In turn, such belief hierarchies have been modelled by *epistemic models* with types, which may be seen as an easy way to *encode* such infinite belief hierarchies.

But if this is true, then we could attempt to formalize a reasoning context as follows: Take an arbitrary dynamic game G and an epistemic model M. Then, a reasoning context selects for a given player a subset of types within M, representing those belief hierarchies that are "allowed for" by the reasoning context. Although this may seem reasonable there is one major problem with this attempt, namely that the epistemic model at hand may not contain *all* belief hierarchies that we are interested in – some belief hierarchies that we would wish to select are simply not present in the epistemic model. In order to avoid this problem we assume the epistemic model to be *belief-complete*[3] (cf. Brandenburger (2003)).

Definition 4.1 (Belief complete epistemic model) *Consider a dynamic game G and an epistemic model $M = (T_i, b_i)_{i \in I}$ for G. The epistemic model M is belief-complete if for every player i, and every possible conditional belief vector $\beta_i = (\beta_i(h))_{h \in H_i}$ for player i, where $\beta_i(h) \in \Delta(S_j(h) \times T_j)$ for every $h \in H_i$, there is some type $t_i \in T_i$ for which $b_i(t_i, h) = \beta_i(h)$ for every $h \in H_i$.*

That is, for every possible conditional belief vector that we can construct there is a type that has precisely this belief vector. It is not at all obvious that such models will always exist. Battigalli and Siniscalchi (1999), however, have shown that for every finite dynamic game, we can always construct a belief-complete epistemic model which assumes (common belief in) Bayesian updating. A similar construction can be employed to build a belief-complete epistemic model without Bayesian updating, as we use here. Formally speaking, there may be various different belief-complete epistemic models for a given dynamic game. However, all such belief-complete epistemic models may be viewed as "equivalent", since each of these encodes all possible conditional belief hierarchies we can think of.

So, if we work with a belief-complete epistemic model, then we are sure not to miss out on any conditional belief vector we could possibly have constructed within our model. With this definition at hand, we can now define a reasoning context as a mapping that selects a subset of belief hierarchies within a *belief-complete* epistemic model.

Definition 4.2 (Reasoning context) *A reasoning context is a mapping ρ that assigns to every finite dynamic game G, every belief-complete epistemic model $M = (T_i, b_i)_{i \in I}$ for G, and every player $i \in I$, some subset $\rho_i(G, M) \subseteq T_i$ of types.*

So, effectively, a reasoning context selects for every dynamic game a set of belief hierarchies for every player. Examples of reasoning contexts for dynamic games are *common belief in future rationality* (Perea (2014), Baltag, Smets and Zvesper (2009), Penta (2009)), *sequential rationalizability* (Dekel, Fudenberg and Levine (1999, 2002), Asheim and Perea (2005)), and *common strong belief in rationality* (Battigalli and Siniscalchi (2002)).

[3] Brandenburger (2003) uses the term *complete.* Following Amanda Friedenberg's suggestion I use the term *belief complete* instead, as it reveals more precisely its content. Moreover, it avoids a possible confusion with the term *complete* as used in complete topological spaces.

4.2 Equilibrium Reasoning

A common feature of all equilibrium concepts for static and dynamic games – such as Nash equilibrium (Nash (1950, 1951)), perfect equilibrium (Selten (1975)), proper equilibrium (Myerson (1978)), subgame perfect equilibrium (Selten (1965)) and sequential equilibrium (Kreps and Wilson (1982)) – is that they require each player to believe that his opponents are correct about the actual beliefs he holds, and that he believes that all other players also believe that their opponents are correct about the actual beliefs they have. In fact, for two-player games this is exactly the condition that separates a rationalizability concept from its equilibrium counterpart, as described in Table 1 in the introduction. See Brandenburger and Dekel (1987, 1989), Tan and Werlang (1988), Aumann and Brandenburger (1995), Asheim (2006) and Perea (2007) for the case of Nash equilibrium[4], see Perea and Predtetchinski (2014) for the case of subgame perfect equilibrium and sequential equilibrium[5], and see Perea (2012, Problems 5.9 and 6.9) for the case of perfect equilibrium and proper equilibrium[6].

Within an epistemic model, the "correct beliefs assumption" can be formalized in the following way.

Definition 4.3 (Equilibrium reasoning) *Consider a dynamic game G and an epistemic model $M = (T_i, b_i)_{i \in I}$ for G.*

(a) For a given type $t_i \in T_i$, let $T_j(t_i)$ be the set of types t_j for opponent j for which $b_j(t_j, h)(S_i \times \{t_i\}) = 1$ for every $h \in H_j$. We say that type $t_i \in T_i$ believes that j is correct about his actual beliefs if $b_i(t_i, h)(S_j \times T_j(t_i)) = 1$ for every $h \in H_i$.

(b) Let T_j^* be the set of types for j that believe that i is correct about j's actual beliefs. We say that type t_i believes that opponent j believes that i is correct about j's actual beliefs if

[4] Note that Nash equilibrium is defined for static games, whereas we are dealing with dynamic games in this paper. However, static games can be viewed as a special case of a dynamic game, where there is only one non-terminal history – the beginning of the game – where all players are active. So, the analysis of Nash equilibrium can be embedded in our framework.

[5] To be precise, Perea and Predtetchinski (2014) focus on the class of stochastic games with finitely many states. Since this class encompasses all finite dynamic games with observable past choices, their analysis applies in particular to the setting in this paper. Perea and Predtetchinski (2014) show, for the class of two-player stochastic games with finitely many states, that the conditions which separate common belief in future rationality from subgame perfect equilibrium and sequential equilibrium are the correct beliefs assumption as stated above, together with the condition that both players satisfy Bayesian updating, and believe the opponent to satisfy Bayesian updating as well.

[6] In Perea (2012), perfect equilibrium, proper equilibrium, and the associated rationalizability counterparts, are defined by means of *lexicographic* beliefs, rather than standard probabilistic beliefs. Hence, formally, the analysis in Perea (2012) falls outside the scope of this paper. However, it can still be shown that if one takes the concept of permissibility (Brandenburger (1992), Börgers (1994)) for two-player games, defines it by means of lexicographic beliefs, and then adds the conditions that each player i believes that j is correct about i's beliefs, and believes that j believes that i is correct about j's beliefs (stated in terms of lexicographic beliefs), then one gets precisely the concept of *perfect equilibrium* (Selten (1975)). A similar relationship holds between *proper rationalizability* (Schuhmacher (1999), Asheim (2001)) and *proper equilibrium* (Myerson (1978)).

$b_i(t_i, h)(S_j \times T_j^*) = 1$ for all $h \in H_i$.

(c) Finally, we say that type t_i exhibits equilibrium reasoning if t_i believes that j is correct about his actual beliefs, and if t_i believes that opponent j believes that i is correct about j's actual beliefs.

Remember that a type in an epistemic model is an encoding of a conditional belief hierarchy. Hence, condition (a) states that player i believes, at every history in H_i, that opponent j is correct, at every history in H_j, about his entire conditional belief hierarchy. In other words, player i believes that j is correct about his actual beliefs – precisely the condition we stated verbally above. Similarly, condition (b) formalizes the statement that i believes, throughout the game, that j believes, throughout the game, that i is always correct about j's belief hierarchy.

For games with more than two players, the correct beliefs assumption would no longer be sufficient to characterize equilibrium reasoning. Additional requirements would be needed, such as the condition that a player believes that an opponent shares his beliefs about a third player, or that the belief about one opponent must be stochastically independent from the belief about another opponent. Various epistemic characterizations of Nash equilibrium for games with *more than two players* have been given in Brandenburger and Dekel (1987), Aumann and Brandenburger (1995), Perea (2007), Barelli (2009) and Bach and Tsakas (2014). But, as we said, in this paper we restrict to two-player games, mainly for the ease of exposition.

4.3 Reasoning Context Consistent with Equilibrium Reasoning

With the above definitions at hand, we are now ready to define what it means for a reasoning context to be consistent with equilibrium reasoning.

Definition 4.4 (Reasoning context consistent with equilibrium reasoning) *We say that a reasoning context ρ is consistent with equilibrium reasoning in a specific dynamic game G, if for every belief-complete epistemic model $M = (T_i, b_i)_{i \in I}$ for G, and every player i, there is some type $t_i \in \rho_i(G, M)$ that exhibits equilibrium reasoning.*

Remember that a reasoning context ρ assigns to every dynamic game G, and every belief-complete epistemic model $M = (T_i, b_i)_{i \in I}$ for G, a subset of types $\rho_i(G, M) \subseteq T_i$ for both players i. The condition above thus states that for the dynamic game G at hand there must be at least one belief hierarchy for each player that exhibits equilibrium reasoning, and that simultaneously satisfies the conditions imposed by the reasoning context ρ. In other words, the conditions imposed by the reasoning context ρ must be logically consistent with the conditions imposed by equilibrium reasoning.

In particular, Table 1 shows that the reasoning contexts of *common belief in rationality, permissibility* and *proper rationalizability* for static games, and the reasoning context of *common belief in future rationality* for dynamic games, are all consistent with equilibrium reasoning in every game. The second column in Table 1 describes, for each of these reasoning contexts, the equilibrium concept one would obtain if we were to add the additional requirement that types must exhibit equilibrium reasoning in every game. In the following section we will show that the reasoning context of *common strong belief in rationality* – which represents a very natural form of forward induction reasoning – is inconsistent with equilibrium reasoning in some (in fact, most) dynamic games.

5 CSBR Can Be Inconsistent with Equilibrium Reasoning

In this section we formally define the reasoning context of *common strong belief in rationality* (Battigalli and Siniscalchi (2002)), argue that it represents a very basic and pure form of forward induction reasoning, and show that it is inconsistent with equilibrium reasoning in some dynamic games. In the next section we will show that the class of dynamic games for which it *is* consistent with equilibrium reasoning is in fact very small.

5.1 Common Strong Belief in Rationality

The reasoning context of "common strong belief in rationality" has been developed by Battigalli and Siniscalchi (2002). They have shown that the strategies that can rationally be chosen by players who reason in accordance with this concept correspond precisely to the *extensive-form rationalizable* strategies as defined by Pearce (1984) and Battigalli (1997). The main idea behind "common strong belief in rationality" is that a player must believe in the opponent's rationality whenever this is possible – a typical forward induction argument. More precisely, if player i finds himself at history $h \in H_i$, and concludes that h *could* be reached if j chooses rationally, then player i *must* believe at h that j chooses rationally. We say that player i *strongly believes* in j's rationality. Moreover, if h could be reached if j chose rationally, then player i asks a second question: could h still be reached if j not only chooses rationally but also strongly believes in i's rationality? If the answer is yes, then player i *must* believe at h that j chooses rationally *and* strongly believes in i's rationality. By iterating this argument, we arrive at "common strong belief in rationality".

In a sense, a player tries to find, at each history where is active, a "best possible explanation" for the past opponent's choices he has observed so far, and uses this explanation to form a belief about the opponent's current and future choices. Common strong belief in rationality can therefore be viewed as a very basic and pure form of forward induction reasoning. To formalize the notion of common strong belief in rationality, let us first define what we mean by *rationality* and *strong belief.*

Consider a type t_i for player i, a history $h \in H_i$ and a strategy s_i that is consistent with h. By $u_i(s_i, b_i(t_i, h))$ we denote the expected utility that player i gets if the game is at h, player i chooses s_i there, and holds the conditional belief $b_i(t_i, h)$ about the opponent's strategy-type combinations. Note that this expected utility does not depend on the full conditional belief that t_i holds at h, but only on the conditional belief about the opponent's strategy choice.

Definition 5.1 (Rational choice) *Consider a type t_i for player i, a history $h \in H_i$ and a strategy s_i that is consistent with h. Strategy s_i is rational for type t_i at history h if $u_i(s_i, b_i(t_i, h)) \geq u_i(s_i', b_i(t_i, h))$ for all alternative strategies s_i' that are consistent with h. Strategy s_i is rational for type t_i if it is so at every history $h \in H_i$ that s_i is consistent with.*

In words, a strategy is rational for a type if at every relevant history it yields the highest possible expected utility, given the conditional belief held by the type at that history. We next define the notion of strong belief.

Definition 5.2 (Strong belief) *Consider a type t_i within a belief-complete epistemic model $M = (T_i, b_i)_{i \in I}$, and an event $E \subseteq S_j \times T_j$. Type t_i strongly believes the event E if $b_i(t_i, h)(E) = 1$ at every history $h \in H_i$ where $(S_j(h) \times T_j) \cap E$ is non-empty.*

That is, at every history $h \in H_i$ where the event E is consistent with the event of h being reached, player i must concentrate his belief fully on E. The reasoning context of "common strong belief in rationality" can now be defined as follows.

Definition 5.3 (Common strong belief in rationality) *Consider a dynamic game G and a belief-complete epistemic model $M = (T_i, b_i)_{i \in I}$. For every player i we recursively define sets T_i^k and R_i^k as follows.*

Induction start. *Define $T_i^0 := T_i$ and $R_i^0 := \{(s_i, t_i) \in S_i \times T_i^0 \mid s_i$ rational for $t_i\}$.*

Induction step. *Let $k \geq 1$, and suppose T_i^{k-1} and R_i^{k-1} have been defined for all players i. Then, for both players i,*

$$\begin{aligned} T_i^k &: = \{t_i \in T_i^{k-1} \mid t_i \text{ strongly believes } R_j^{k-1}\}, \text{ and} \\ R_i^k &: = \{(s_i, t_i) \in S_i \times T_i^k \mid s_i \text{ rational for } t_i\}. \end{aligned}$$

Common strong belief in rationality selects for every player i the set of types $T_i^\infty := \cap_{k \in \mathbb{N}} T_i^k$.

We say that a type t_i expresses *common strong belief in rationality* if $t_i \in T_i^\infty$. Battigalli and Siniscalchi (2002) show that the sets of types T_i^∞ are always non-empty for every finite dynamic game, and that the strategies which are rational for a type in T_i^∞ are precisely the *extensive-form rationalizable* strategies as defined in Pearce (1984) and Battigalli (1997).

5.2 Example Revisited

Consider again the dynamic game G from Figure 1. We will now formally show that the reasoning context of *common strong belief in rationality* is inconsistent with equilibrium reasoning in the game G. To that purpose, take an arbitrary belief-complete epistemic model $M = (T_i, b_i)_{i \in I}$ for G, and let $T_i^\infty \subseteq T_i$ be the set of types selected by common strong belief in rationality, for both players i. We will show that there is no type $t_1 \in T_1^\infty$ that exhibits equilibrium reasoning.

Take an arbitrary type $t_1^* \in T_1^\infty$. Then, in particular, t_1^* strongly believes R_2^1. This implies that $b_1(t_1^*, \emptyset)(R_2^1) = 1$, since $(S_2(\emptyset) \times T_2) \cap R_2^1 = (S_2 \times T_2) \cap R_2^1 \neq \emptyset$. As R_2^1 only contains strategy-type pairs (s_2, t_2) where s_2 is rational for t_2 at h_1, it follows that $R_2^1 \subseteq \{e, f\} \times T_2$. Hence,

$$b_1(t_1^*, \emptyset)(\{e, f\} \times T_2) = 1. \tag{1}$$

Now, take an arbitrary type $t_2 \in T_2^1$. Then, t_2 strongly believes R_1^0. Since the epistemic model M is belief-complete, there is a type $t_1 \in T_1$ with $b_1(t_1, \emptyset)(\{g\} \times T_2) = b_1(t_1, h_1)(\{g\} \times T_2) = 1$. Since $((a, d), t_1) \in (S_1(h_1) \times T_1) \cap R_1^0$, it follows that $(S_1(h_1) \times T_1) \cap R_1^0 \neq \emptyset$. But then, as t_2 strongly believes R_1^0, we conclude that $b_2(t_2, h_1)(R_1^0) = 1$.

Since, by equation (1), type t_1^* assigns probability 0 to choice g by player 2, there is no strategy $s_1 \in S_1(h_1)$ for which $(s_1, t_1^*) \in R_1^0$. To see this, note that $S_1(h_1) = \{(a, c), (a, d)\}$, and both (a, c) and (a, d) yield t_1^* an expected utility less than 3 – the utility it can guarantee at $\emptyset$ by choosing b. As $b_2(t_2, h_1)((S_1(h_1) \times T_1) \cap R_1^0) = 1$, this implies that $b_2(t_2, h_1)(S_1 \times \{t_1^*\}) = 0$. Hence, we see that

$$b_2(t_2, h_1)(S_1 \times \{t_1^*\}) = 0 \text{ for every type } t_2 \in T_2^1. \tag{2}$$

We have seen above that $b_1(t_1^*, \emptyset)(R_2^1) = 1$, and hence, in particular,

$$b_1(t_1^*, \emptyset)(S_2 \times T_2^1) = 1, \tag{3}$$

since $R_2^1 \subseteq S_2 \times T_2^1$.

By combining (2) and (3) we see that t_1^* assigns at $\emptyset$ probability 1 to the set of types T_2^1, but that every type $t_2 \in T_2^1$ assigns, at h_1, probability 0 to player 1's type t_1^*. This means, however, that type t_1^* cannot exhibit equilibrium reasoning.

Since this holds for every type $t_1^* \in T_1^\infty$, we conclude that there is no type $t_1^* \in T_1^\infty$ that exhibits equilibrium reasoning. As such, the reasoning context of *common strong belief in rationality* is inconsistent with equilibrium reasoning in the game G from Figure 1. We therefore obtain the following result.

Theorem 5.4 (Inconsistency with equilibrium reasoning) *There are two-player dynamic games with observable past choices for which common strong belief in rationality is inconsistent with equilibrium reasoning.*

This formally proves the entry in Table 1 which states that there is no equilibrium counterpart to the reasoning context of common strong belief in rationality. The theorem above immediately raises the following question: Can we characterize those games for which common strong belief in rationality *is* consistent with equilibrium reasoning? That will be the purpose of the following section.

6 When CSBR is Consistent with Equilibrium Reasoning

In this section we will characterize the class of dynamic games for which common strong belief in rationality is consistent with equilibrium reasoning. For this characterization we will introduce a new concept called *extensive-form best response set with unique beliefs* – a refinement of the notion of *extensive-form best response set* as defined by Battigalli and Friedenberg (2012).

6.1 Extensive-form Best Response Set with Unique Beliefs

In order to formally introduce extensive-form best response sets with unique beliefs, we need the following definitions. A *conditional belief vector* for player i is a tuple $b_i = (b_i(h))_{h \in H_i}$ which assigns to every history $h \in H_i$ a probabilistic belief $b_i(h) \in \Delta(S_j(h))$ on the opponent's strategy choices that are consistent with h. Hence, the first-order belief of a type in some epistemic model is a conditional belief vector in this sense. We say that the conditional belief vector b_i *strongly believes* an event $D_j \subseteq S_j$ if $b_i(h)(D_j) = 1$ at every history $h \in H_i$ where $S_j(h) \cap D_j \neq \emptyset$. Finally, a strategy s_i is said to be *rational* for the conditional belief vector b_i if

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

at every history $h \in H_i$ that s_i is consistent with.

Definition 6.1 (Extensive-form best response set with unique beliefs) *A set $D_1 \times D_2 \subseteq S_1 \times S_2$ of strategy pairs is called an extensive-form best response set with unique beliefs, if for both players i there is a conditional belief vector b_i such that for all strategies $s_i \in D_i$*

(a) s_i is rational for b_i,

(b) b_i strongly believes D_j, and

(c) every strategy s_i' which is rational for b_i is in D_i.

An *extensive-form best response set* as defined in Battigalli and Friedenberg (2012) is a pair $D_1 \times D_2 \subseteq S_1 \times S_2$ such that for both players i and all strategies $s_i \in D_i$ there is a belief vector b_i that satisfies conditions (a), (b) and (c) above. Hence, our notion of an extensive-form best response set with unique beliefs is a special case of an extensive-form best response set à la Battigalli and Friedenberg. The difference is that in the former we choose a *unique* conditional belief vector b_i that satisfies the conditions (a), (b) and (c) for every strategy $s_i \in D_i$, whereas in the latter we may choose a *different* conditional belief vector b_i for every strategy $s_i \in D_i$ we consider.

6.2 When is CSBR Consistent with Equilibrium Reasoning?

We will now characterize the class of dynamic games for which common strong belief in rationality is consistent with equilibrium reasoning. Before we state our result, we need a few more definitions. We call a strategy $s_i \in S_i$ *rational* if there is a conditional belief vector b_i for which s_i is rational. Say that a history $h \in H_i$ is *consistent with* j*'s rationality* if there is a rational strategy s_j for opponent j that is consistent with h.

Theorem 6.2 (When CSBR is consistent with equilibrium reasoning) *Consider a finite two-player dynamic game G with observed past choices. Then, common strong belief in rationality is consistent with equilibrium reasoning at G, if and only if, there is an extensive-form best response set $D_1 \times D_2$ with unique beliefs such that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j$ that is consistent with h.*

The proof of this theorem can be found in Section 8. This theorem implies that in "most" two-player dynamic games with observable past choices, common strong belief in rationality is *inconsistent* with equilibrium reasoning. Indeed, assume that in the game G common strong belief in rationality is consistent with equilibrium reasoning. Then, according to Theorem 6.2, there is an extensive-form best response set $D_1 \times D_2$ with unique beliefs such that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j$ that is consistent with h. Hence, there must be a conditional belief vector b_i for both players i such that (1) D_i is the set of rational strategies for b_i, (2) b_i strongly believes D_j, and (3) for every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j$ that is consistent with h. In particular, for both players i there must be a *unique* conditional belief vector b_i such that *every* history $h \in H_j$ that is consistent with i's rationality must be reachable by a strategy s_i that is rational for b_i. This, however, will typically not be the case, as the set of strategies that are rational for a fixed conditional belief vector b_i is typically very small, whereas the collection of histories $h \in H_j$ that is consistent with i's rationality is typically very large, especially when the game G is not too small. We may therefore conclude that "typically", common strong belief in rationality will be *inconsistent* with equilibrium reasoning.

7 Forward Induction Equilibrium Concepts

In Sections 5 and 6 we have shown that the reasoning context of *common strong belief in rationality* is inconsistent with equilibrium reasoning in "most" dynamic games. At the same time, the literature offers a broad spectrum of forward induction *equilibrium* concepts which – by construction – *are* consistent with equilibrium reasoning, and which incorporate some particular form of forward induction. In this section we will show, however, that none of these forward induction equilibrium concepts can single out the intuitive forward induction choice f of player

2 in the game of Figure 1. The reason for this is that each of these concepts imposes some *exogenous restrictions* on the players' reasoning which may substantially obscure, or weaken, the forward induction reasoning.

Let us go back to the dynamic game in Figure 1. Why is f the intuitive forward induction choice of player 2 here? If player 2 must make a choice at h_1 he knows that player 1 has chosen a, and not b, at the beginning. But choosing a can only be optimal for player 1 at the beginning if he subsequently chooses d, and believes that player 2, with high probability, will make the irrational choice g. Hence, player 2 must conclude that player 1 will subsequently choose d. As such, the only natural forward induction choice for player 2 is to choose f.

Note that, in order for player 2 to carry out this forward induction reasoning, he must consider the possibility that player 1 may ascribe a high probability to his *irrational* choice g. We will see that this is exactly where the forward induction *equilibrium* concepts fall short: in each of the equilibrium concepts player 2 does not even consider the possibility that player 1 may assign a positive probability to 2's irrational choice g, and therefore none of these equilibrium concepts is able to uniquely select the intuitive forward induction choice f of player 2.

Let us now be more precise about these claims. The forward induction equilibrium concepts I am aware of consist of *justifiable sequential equilibrium* (McLennan (1985)), *Cho's forward induction equilibrium* (Cho (1987)), *stable sets of beliefs* (Hillas (1994)), *explicable equilibrium* (Reny (1992)), *outcomes satisfying forward induction* (Govindan and Wilson (2009)) and *Man's forward induction equilibrium* (Man (2012)).[7] Of these concepts, the former three are formulated as refinements of *sequential equilibrium* (Kreps and Wilson (1982))[8], the fourth and fifth are defined as refinements of *weak sequential equilibrium* (Reny (1992))[9], whereas the last is a refinement of *normal-form perfect equilibrium* (Selten (1975)).

It may be verified that the game of Figure 1 has a unique sequential equilibrium, in which player 2 chooses e. Consequently, *justifiable sequential equilibrium, Cho's forward induction equilibrium* and *stable sets of beliefs* – being refinements of sequential equilibrium – will uniquely select the choice e for player 2, which is *not* the intuitive forward induction choice, as we have seen.

In every *weak* sequential equilibrium of the game in Figure 1, player 1 assigns probability 0 to player 2 choosing g. As the reasoning of players in *explicable equilibrium* and *outcomes satisfying forward induction* takes place entirely within the framework of weak sequential equilibria, player 2 – in these concepts – cannot even reason about the possibility that player 1 assigns a positive

[7] There are other forward induction equilibrium concepts that are only applicable to signaling games, such as the *intuitive criterion* (Cho and Kreps (1987)), *divine equilibrium* (Banks and Sobel (1987)), *perfect sequential equilibrium* (Grossman and Perry (1986)) and *undefeated equilibrium* (Mailath, Okuno-Fujiawara and Postlewaite (1993)).

[8] Strictly speaking, Hillas' concept of *stable sets of beliefs* imposes restrictions on *sets* of sequential equilibria, rather than on *individual* sequential equilibria. However, it *indirectly* imposes restrictions on an individual sequential equilibrium if we require it to be part of a set of sequential equilibria meeting Hillas' conditions.

[9] More precisely, Govindan and Wilson (2009) impose restrictions on *outcomes* rather than weak sequential equilibria. However, these restrictions may be translated into restrictions on weak sequential equilibria directly.

probability to choice g. As such, in these concepts player 2 cannot give a plausible explanation at h_1 for the event that player 1 has chosen a and not b. Consequently, both *explicable equilibrium* and *outcomes satisfying forward induction* impose no restrictions on what player 2 believes at h_1 about 1's choice at h_1. In particular, both *explicable equilibrium* and *outcomes satisfying forward induction* allow for the choices e and f by player 2, and therefore fail to single out the intuitive forward induction choice f for player 2.

Since every normal-form perfect equilibrium induces a weak sequential equilibrium (see Reny (1992), Proposition 1), player 1 must also assign probability 0 to player 2's irrational choice g in every normal-form perfect equilibrium. As the reasoning of players in *Man's forward induction equilibrium* is restricted to the setting of normal-form perfect equilibria, it follows with the same arguments as above that also *Man's forward induction equilibrium* allows for the choices e and f by player 2, and thus fails to uniquely select the intuitive forward induction choice f for player 2.

We thus conclude that none of the forward induction equilibrium concepts above uniquely selects the intuitive forward induction choice f of player 2. What prevents these forward induction equilibrium concepts from uniquely selecting the intuitive forward induction choice f for player 2 is that each of these concepts imposes some *exogenous restrictions* on the beliefs of the players which interfere with – and in some situations are in conflict with – forward induction reasoning. For instance, the first three concepts impose, as an exogenous restriction, that players reason in accordance with sequential equilibrium, which is a *backward induction* concept. That is, the first three concepts assume that players – above all – reason in accordance with backward induction, and on top of this impose some forward induction restrictions. As a result we obtain concepts that are a mix of backward induction and forward induction arguments. As backward induction reasoning alone already singles out the backward induction choice e by player 2, the forward induction arguments in the first three concepts have no bite in the game of Figure 1, and still uniquely lead to the backward induction choice e for player 2.

The last three concepts impose, as an exogenous restriction, that player 1 will always assign probability 0 to 2's irrational choice g. As player 2, under these circumstances, cannot give a rational explanation at h_1 for player 1 choosing a, the forward induction arguments in the last three concepts have no bite either in the game of Figure 1.

In contrast, the "pure" forward induction concept of *common strong belief in rationality* imposes no exogenous restrictions on the beliefs of the players, and therefore allows player 2 to reason about a scenario in which player 1 assigns a high probability to player 2's irrational choice g. This, eventually, makes it possible for common strong belief in rationality to uniquely filter the intuitive forward induction choice f for player 2.

Recently, Battigalli and Friedenberg (2012) have started to study variants of the concept of *common strong belief in rationality* in which they *do* impose exogenous restrictions on the beliefs of the players. To achieve this, Battigalli and Friedenberg use epistemic models that are *not necessarily belief-complete.* In other words, the epistemic model may not contain all possible belief hierarchies for the players. As players can only hold belief hierarchies within the epistemic model, and can only reason about opponent's belief hierarchies that are within that same epistemic model, choosing a belief-incomplete epistemic model imposes some exogenous restrictions on the players' belief hierarchies, which may have drastic consequences for the type of forward induction reasoning they use.

In the game of Figure 1, such an exogenous restriction could be that we only allow for types of player 1 that assign probability 0 to 2's irrational choice g. With such an exogenous restriction, *common strong belief in rationality* loses all of its bite in the game of Figure 1, as player 2, at h_1, can no longer rationalize the event that player 1 has chosen a. Consequently, common strong belief in rationality would allow player 2 to choose either e or f, and hence would no longer uniquely select the intuitive forward induction choice for player 2.

It would be interesting to see whether some of the forward induction equilibrium concepts above can be characterized in the spirit of Battigalli and Friedenberg (2012) by common strong belief in rationality relative to a suitably restricted epistemic model. We leave this question for future research.

8 Proof

Proof of Theorem 6.2. Let $M = (T_i, b_i)_{i \in I}$ be an arbitrary belief-complete epistemic model for G, and let T_i^∞ be the set of types in T_i that express common strong belief in rationality, for both players i.

(a) Suppose first that common strong belief in rationality is consistent with equilibrium reasoning at G. We will show that there is an extensive-form best response set $D_1 \times D_2$ with unique beliefs such that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j$ that is consistent with h.

As common strong belief in rationality is consistent with equilibrium reasoning at G, there is for both players i a type $t_i^* \in T_i^\infty$ that exhibits equilibrium reasoning. Let i be a player who is active at $\emptyset$ – the beginning of the game.

Claim. There is a unique type $t_j^* \in T_j^\infty$ such that $b_i(t_i^*, h)(S_j \times \{t_j^*\}) = 1$ for all $h \in H_i$, and $b_j(t_j^*, h)(S_i \times \{t_i^*\}) = 1$ for all $h \in H_j$.

Proof of claim. Since t_i^* exhibits equilibrium reasoning, type t_i^* believes that j is correct about his actual beliefs. That is, $b_i(t_i^*, h)(S_j \times T_j(t_i^*)) = 1$ for every $h \in H_i$, where

$$T_j(t_i^*) = \{t_j \in T_j \mid b_j(t_j, h)(S_i(h) \times \{t_i^*\}) = 1 \text{ for all } h \in H_j\}.$$

We first show that there is a single type $t_j^* \in T_j(t_i^*)$ such that $b_i(t_i^*, h)(S_j \times \{t_j^*\}) = 1$ for all $h \in H_i$. Suppose not. Then, there are two disjoint sets $\bar{T}_j, \tilde{T}_j \subseteq T_j(t_i^*)$ with $\bar{T}_j \cup \tilde{T}_j = T_j(t_i^*)$ such that $b_i(t_i^*, h^1)(S_j \times \bar{T}_j) > 0$ for some $h^1 \in H_i$, and $b_i(t_i^*, h^2)(S_j \times \tilde{T}_j) > 0$ for some $h^2 \in H_i$.

Take an arbitrary type $t_j \in T_j(t_i^*)$ and assume, with loss of generality, that $t_j \in \bar{T}_j$. As $t_j \in T_j(t_i^*)$ we know that

$$b_j(t_j, h)(S_i \times \{t_i^*\}) = 1 \text{ for all } h \in H_j. \tag{4}$$

Moreover, we know from above that $b_i(t_i^*, h^2)(S_j \times \tilde{T}_j) > 0$. Since $t_j \notin \tilde{T}_j$, it follows that

$$b_i(t_i^*, h^2)(S_j \times \{t_j\}) < 1. \tag{5}$$

From (4) and (5) we conclude that type t_j does not believe that i is correct about j's actual beliefs. Since this holds for every $t_j \in T_j(t_i^*)$, and $b_i(t_i^*, h)(S_j \times T_j(t_i^*)) = 1$ for every $h \in H_i$, it follows that t_i^* does *not* believe that j believes that i is correct about j's actual beliefs. This, however, is a contradiction, since t_i^* exhibits equilibrium reasoning. We may thus conclude that there is some $t_j^* \in T_j(t_i^*)$ such that $b_i(t_i^*, h)(S_j \times \{t_j^*\}) = 1$ for all $h \in H_i$.

Since $t_j^* \in T_j(t_i^*)$, it immediately follows that $b_j(t_j^*, h)(S_i \times \{t_i^*\}) = 1$ for all $h \in H_j$. Moreover, as player i is active at $\emptyset$ and $t_i^* \in T_i^\infty$ expresses common strong belief in rationality, it follows that $b_i(t_i^*, \emptyset)(S_j \times T_j^\infty) = 1$, which implies that $t_j^* \in T_j^\infty$.

Hence, there is a unique type $t_j^* \in T_j^\infty$ such that $b_i(t_i^*, h)(S_j \times \{t_j^*\}) = 1$ for all $h \in H_i$, and $b_j(t_j^*, h)(S_i \times \{t_i^*\}) = 1$ for all $h \in H_j$, which completes the proof of the claim.

Now, let D_i^* be the set of strategies that are rational for t_i^*, and let D_j^* be the set of strategies that are rational for t_j^*. We show that $D_i^* \times D_j^*$ is an extensive-form best response set with unique beliefs such that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j$ that is consistent with h.

We first prove that $D_i^* \times D_j^*$ is an extensive-form best response set with unique beliefs. Let b_i^* be the first-order conditional belief vector of type t_i^*, and b_j^* the first-order conditional belief vector of type t_j^*. Then, by construction, D_i^* is the set of strategies that are rational for b_i^*. Moreover, as t_i^* expresses common strong belief in rationality, we have in particular that t_i^* strongly believes

$$R_j^0 = \{(s_j, t_j) \in S_j \times T_j \mid s_j \text{ rational for } t_j\}.$$

Together with the fact that $b_i(t_i^*, h)(S_j \times \{t_j^*\}) = 1$ for all $h \in H_i$, this implies that t_i^* strongly believes the event

$$\begin{aligned} R_j^0 \cap (S_j \times \{t_j^*\}) &= \{(s_j, t_j^*) \in S_j \times \{t_j^*\} \mid s_j \text{ rational for } t_j^*\} \\ &= D_j^* \times \{t_j^*\}. \end{aligned}$$

Hence, t_i^*'s first-order conditional belief vector b_i^* strongly believes D_j^*. Summarizing, we see that D_i^* is the set of strategies that are rational for b_i^*, and that b_i^* strongly believes D_j^*. As the same applies to D_j^* and b_j^*, we may conclude that $D_i^* \times D_j^*$ is an extensive-form best response set with unique beliefs.

We finally show that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j^*$ that is consistent with h. For an arbitrary player i, take an arbitrary history $h \in H_i$ that is consistent with j's rationality. We must prove that there is some $s_j \in D_j^*$ that is consistent with h.

As history h is consistent with j's rationality, we have that

$$R_j^0 \cap (S_j(h) \times T_j) \neq \emptyset.$$

Since t_i^* expresses common strong belief in rationality, it strongly believes R_j^0, and hence

$$b_i(t_i^*, h)(R_j^0) = 1.$$

Moreover, as $b_i(t_i^*, h)(S_j \times \{t_j^*\}) = 1$, it follows that

$$\begin{aligned} & b_i(t_i^*, h)(R_j^0 \cap (S_j(h) \times \{t_j^*\})) \\ = \quad & b_i(t_i^*, h)(\{(s_j, t_j^*) \in S_j \times \{t_j^*\} \mid s_j \text{ rational for } t_j^*\}) \\ = \quad & b_i(t_i^*, h)(D_j^* \times \{t_j^*\}) = 1, \end{aligned}$$

which implies that there must be a strategy $s_j \in D_j^*$ that is consistent with h. We thus conclude that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j^*$ that is consistent with h.

We have therefore shown that $D_i^* \times D_j^*$ is an extensive-form best response set with unique beliefs such that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j^*$ that is consistent with h. This concludes the proof of part (a).

(b) Suppose now that there is an extensive-form best response set $D_1 \times D_2$ with unique beliefs such that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j$ that is consistent with h. We show that common strong belief in rationality is consistent with equilibrium reasoning at G.

Take an extensive-form best response set $D_1^* \times D_2^*$ with unique beliefs such that for every player i and every history $h \in H_i$ that is consistent with j's rationality there is a strategy $s_j \in D_j^*$ that is consistent with h. Then, by definition, there is for both players i a conditional belief vector b_i^* such that D_i^* is the set of strategies that are rational for b_i^*, and b_i^* strongly believes D_j^*. Let $t_1^* \in T_1$ and $t_2^* \in T_2$ be types such that for both players i,

$$b_i(t_i^*, h)(\{(s_j, t_j^*)\}) := b_i^*(h)(s_j) \tag{6}$$

for all histories $h \in H_i$ and all $s_j \in S_j(h)$. As $M = (T_i, b_i)_{i \in I}$ is a belief complete epistemic model, such types t_1^* and t_2^* exist. By (6) it immediately follows that

$$b_1(t_1^*, h)(S_2 \times \{t_2^*\}) = 1 \text{ for all } h \in H_1,$$

and

$$b_2(t_2^*, h)(S_1 \times \{t_1^*\}) = 1 \text{ for all } h \in H_2,$$

and hence both types t_1^* and t_2^* exhibit equilibrium reasoning.

We will now show that t_1^* and t_2^* also express common strong belief in rationality. To that purpose we prove, by induction on k, that $t_1^* \in T_1^k$ and $t_2^* \in T_2^k$ for all $k \geq 0$.

For $k = 0$ the statement is trivial, since $T_i^0 = T_i$ for both players i.

Take now some $k \geq 1$, and assume that $t_i^* \in T_i^{k-1}$ for both players i. Choose a player i. In order to show that $t_i^* \in T_i^k$, it only remains to prove that t_i^* strongly believes R_j^{k-1}, as $t_i^* \in T_i^{k-1}$ by the induction assumption.

Consider a history $h \in H_i$ where $R_j^{k-1} \cap (S_j(h) \times T_j) \neq \emptyset$. Then, in particular, h is consistent with j's rationality. By our assumption above, there is a strategy $s_j \in D_j^*$ that is consistent with h. That is, $S_j(h) \cap D_j^* \neq \emptyset$. As b_i^* strongly believes D_j^*, it follows that $b_i^*(h)(D_j^*) = 1$. But then, by (6) it follows that

$$b_i(t_i^*, h)(D_j^* \times \{t_j^*\}) = 1. \tag{7}$$

By construction, D_j^* is the set of rational strategies for b_j^*, and hence also the set of rational strategies for t_j^*, since b_j^* is the first-order conditional belief vector of t_j^*. Since, by the induction assumption, $t_j^* \in T_j^{k-1}$, it follows that

$$D_j^* \times \{t_j^*\} \subseteq R_j^{k-1}. \tag{8}$$

If we combine (7) and (8), we obtain that

$$b_i(t_i^*, h)(R_j^{k-1}) = 1.$$

Hence, we have shown that $b_i(t_i^*, h)(R_j^{k-1}) = 1$ for every $h \in H_i$ where $R_j^{k-1} \cap (S_j(h) \times T_j) \neq \emptyset$, which means that t_i^* strongly believes R_j^{k-1}. Hence, by definition, it follows that $t_i^* \in T_i^k$. As this holds for both players i, it follows by induction that $t_1^* \in T_1^\infty$ and $t_2^* \in T_2^\infty$, as was to show.

Overall, we see that there are types $t_1^* \in T_1^\infty$ and $t_2^* \in T_2^\infty$ that exhibit equilibrium reasoning. Hence, common strong belief in rationality is consistent with equilibrium reasoning at G. This completes the proof of the theorem. ■

References

[1] Asheim, G.B. (2001), Proper rationalizability in lexicographic beliefs, *International Journal of Game Theory* **30,** 453–478.

[2] Asheim, G.B. (2006), *The consistent preferences approach to deductive reasoning in games,* Theory and Decision Library, Springer, Dordrecht, The Netherlands.

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

[4] Aumann, R. and A. Brandenburger (1995), Epistemic conditions for Nash equilibrium, *Econometrica* **63,** 1161–1180.

[5] Bach, C.W. and E. Tsakas (2014), Pairwise epistemic conditions for Nash equilibrium, *Games and Economic Behavior* **85,** 48–59.

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

[7] Banks, J.S. and J. Sobel (1987), Equilibrium selection in signaling games, *Econometrica* **55,** 647–661.

[8] Barelli, P. (2009), Consistency of beliefs and epistemic conditions for Nash and correlated equilibria, *Games and Economic Behavior* **67,** 363–375.

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

[10] Battigalli, P. and A. Friedenberg (2012), Forward induction reasoning revisited, *Theoretical Economics* **7,** 57–98.

[11] Battigalli, P. and M. Siniscalchi (1999), Hierarchies of conditional beliefs and interactive epistemology in dynamic games, *Journal of Economic Theory* **88**, 188–230.

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

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

[14] Börgers, T. (1994), Weak dominance and approximate common knowledge, *Journal of Economic Theory* **64**, 265–276.

[15] Brandenburger, A. (1992), Lexicographic probabilities and iterated admissibility, in P. Dasgupta *et al.* (eds.), *Economic Analysis of Markets and Games* (MIT Press, Cambridge, MA), pp. 282–290.

[16] Brandenburger, A. (2003), On the existence of a "complete" possibility structure, in N. Dimitri, M. Basili and I. Gilboa (eds.), *Cognitive Processes and Economic Behavior* (Routledge, London).

[17] Brandenburger, A. and E. Dekel (1987), Rationalizability and correlated equilibria, *Econometrica* **55,** 1391–1402.

[18] Brandenburger, A. and E. Dekel (1989), The role of common knowledge assumptions in game theory, in *The Economics of Missing Markets, Information and Games,* ed. by Frank Hahn. Oxford: Oxford University Press, pp. 46–61.

[19] Brandenburger, A., Friedenberg, A. and J. Keisler (2008), Admissibility in games, *Econometrica* **76**, 307–352.

[20] Cho, I.-K. (1987), A refinement of sequential equilibrium, *Econometrica* **55,** 1367–1389.

[21] Cho, I.-K., and D.M. Kreps (1987), Signaling games and stable equilibria, *Quarterly Journal of Economics* **102,** 179–221.

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

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

[24] Govindan, S. and R. Wilson (2009), On forward induction, *Econometrica* **77,** 1–28.

[25] Grossman, S.J. and M. Perry (1986), Perfect sequential equilibrium, *Journal of Economic Theory* **39,** 97–119.

[26] Harsanyi, J.C. (1967–1968), Games with incomplete information played by “bayesian” players, I–III’, *Management Science* **14**, 159–182, 320–334, 486–502.

[27] Hillas, J. (1994), Sequential equilibria and stable sets of beliefs, *Journal of Economic Theory* **64,** 78–102.

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

[29] Mailath, G.J., Okuno-Fujiwara, M. and A. Postlewaite (1993), Belief-based refinements in signalling games, *Journal of Economic Theory* **60,** 241–276.

[30] Man, P.T.Y. (2012), Forward induction equilibrium, *Games and Economic Behavior* **75,** 265–276.

[31] McLennan, A. (1985), Justifiable beliefs in sequential equilibria, *Econometrica* **53,** 889–904.

[32] Myerson, R.B. (1978), Refinements of the Nash equilibrium concept, *International Journal of Game Theory* **7**, 73–80.

[33] Nash, J.F. (1950), Equilibrium points in N-person games, *Proceedings of the National Academy of Sciences of the United States of America* **36**, 48–49.

[34] Nash, J.F. (1951), Non-cooperative games, *Annals of Mathematics* **54**, 286–295.

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

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

[37] Perea, A. (2007), A one-person doxastic characterization of Nash strategies, *Synthese* **158**, 251–271 (*Knowledge, Rationality and Action* 341–361).

[38] Perea, A. (2012), *Epistemic Game Theory: Reasoning and Choice,* Cambridge University Press.

[39] Perea, A. (2014), Belief in the opponents' future rationality, *Games and Economic Behavior* **83,** 231–254.

[40] Perea, A. and A. Predtetchinski (2014), An epistemic approach to stochastic games, *EPICENTER* Working Paper No. 3,
http://epicenter.name/Perea/Papers/StochasticGames.pdf

[41] Polak, B. (1999), Epistemic conditions for Nash equilibrium, and common knowledge of rationality, *Econometrica* **67,** 673–676.

[42] Reny, P.J. (1992), Backward induction, normal form perfection and explicable equilibria, *Econometrica* **60,** 627–649.

[43] Rubinstein, A. (1991), Comments on the interpretation of game theory, *Econometrica* **59**, 909–924.

[44] Schuhmacher, F. (1999), Proper rationalizability and backward induction, *International Journal of Game Theory* **28**, 599–615.

[45] Selten, R. (1965), Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragezeit, *Zeitschrift für die Gesammte Staatswissenschaft* **121,** 301–324, 667–689.

[46] Selten, R. (1975), Reexamination of the perfectness concept for equilibrium points in extensive games, *International Journal of Game Theory* **4,** 25–55.

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