Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 169 (2017) 489–516

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Forward induction reasoning and correct beliefs ☆

Andrés Perea

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

Received 30 March 2016; final version received 14 February 2017; accepted 27 February 2017
Available online 6 March 2017

Abstract

All equilibrium concepts implicitly make a correct beliefs assumption, stating that a player believes that his opponents are correct about his first-order beliefs. In this paper we show that in many dynamic games of interest, this correct beliefs assumption may be incompatible with a very basic form of forward induction reasoning: the first two layers of extensive-form rationalizability (Pearce, 1984; Battigalli, 1997, epistemically characterized by Battigalli and Siniscalchi, 2002). Hence, forward induction reasoning naturally leads us away from equilibrium reasoning. In the second part we classify the games for which equilibrium reasoning *is* consistent with this type of forward induction reasoning, and find that this class is very small.

JEL classification: C72

Keywords: Epistemic game theory; Dynamic games; Forward induction; Common strong belief in rationality; Correct beliefs assumption; Equilibrium

☆ I thank Christian Bach, Michael Greinecker, Marciano Siniscalchi (in his role as editor) and some anonymous referees for their very valuable comments on this paper. I am also grateful to the audiences at Turku University, University of Innsbruck, Maastricht University, University of Groningen, London School of Economics, the University of Toulouse, LOFT 2016 and GAMES 2016 for their feedback. This paper previously appeared with the title “Forward induction reasoning versus equilibrium reasoning”.

E-mail address: a.perea@maastrichtuniversity.nl.
URL: http://www.epicenter.name/Perea/.

http://dx.doi.org/10.1016/j.jet.2017.02.013
0022-0531/© 2017 Elsevier Inc. All rights reserved.

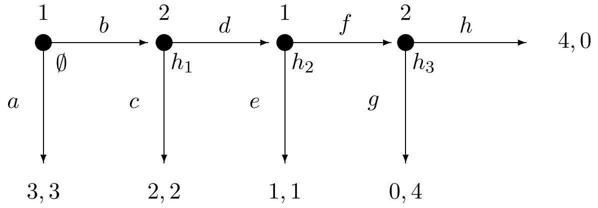


Fig. 1. Reny's game.

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 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.

But what precisely is it that distinguishes rationalizability concepts from equilibrium concepts? To answer that question we must explicitly investigate the first-order and higher-order beliefs[1] of the players, which leads us to the field of epistemic game theory. Several papers in that literature show that equilibrium concepts make a *correct beliefs assumption*, stating that a player believes that his opponents are correct about his first-order belief, whereas rationalizability concepts do not make this assumption. For the case of Nash equilibrium this has been shown in Brandenburger and Dekel (1987, 1989), Tan and Werlang (1988), Aumann and Brandenburger (1995), Asheim (2006) and Perea (2007), which all provide epistemic characterizations of Nash equilibrium that involve, in some way, the correct beliefs assumption above. In a similar fashion, epistemic characterizations of other equilibrium concepts, like perfect equilibrium (Selten, 1975), proper equilibrium (Myerson, 1978), subgame perfect equilibrium (Selten, 1965) and sequential equilibrium (Kreps and Wilson, 1982), also rely on the correct beliefs assumption. In that light, the correct beliefs assumption may be viewed as the essential ingredient of equilibrium reasoning.

The main message of this paper is to show that within the class of dynamic games, the correct beliefs assumption, and hence equilibrium reasoning, is incompatible with a very basic form of forward induction reasoning. Therefore, in order to implement this type of forward induction reasoning we must necessarily leave the context of equilibrium reasoning. As an illustration of this fact, consider the game in Fig. 1, which is based on Fig. 3 in Reny (1992a). It is natural to assume that player 1, at the beginning of the game, believes that player 2 will not choose h at history h_3. Suppose now that player 2, at history h_1, observes that player 1 has chosen b. Since choice b can only be optimal for player 1 if he assigns a high probability to player 2 choosing h, forward induction reasoning seems to suggest that player 2, at h_1, believes that player 1 assigns a high probability to player 2 choosing h. Player 1, anticipating on this type of forward induction reasoning by player 2, therefore believes that player 2, at h_1, will be wrong about his actual first-order belief, thus violating the correct beliefs assumption.

[1] By a first-order belief we mean a belief about the opponents' choices. A second-order belief is a belief about the opponents' choices and first-order beliefs, whereas higher-order beliefs can be defined in a similar fashion.

What we have been using in this example are only the first two layers of extensive-form rationalizability – a very basic and natural forward induction concept developed by Pearce (1984) and later simplified by Battigalli (1997). The first layer states that a player, whenever possible, must believe that his opponents are implementing rational strategies, whereas the second layer requires a player to believe, whenever possible, that his opponents do not only choose rationally but also follow the reasoning of the first layer. Battigalli and Siniscalchi (2002) call the first condition *strong belief in the opponents' rationality,* and it provides the basis for the epistemic condition of *common strong belief in rationality* which, as is shown in Battigalli and Siniscalchi (2002), characterizes extensive-form rationalizability. The example above thus shows that the correct beliefs assumption, underlying equilibrium reasoning, is in conflict with the first two layers of a very basic form of forward induction reasoning. In the first part of this paper we formalize and prove this statement in a precise language.

In the second part we ask for which games equilibrium reasoning *is* compatible with forward induction reasoning. In order to formally address this question we must specify what we mean by "equilibrium reasoning" and "forward induction reasoning". Perea (2007) has shown that in two-player games, Nash equilibrium can be characterized by (the first two layers of) common belief in rationality (Brandenburger and Dekel, 1987; Tan and Werlang, 1988) and a *strong correct beliefs assumption,* stating that player i believes that j is correct about his entire belief hierarchy, and that player i believes that j believes that i is correct about j's entire belief hierarchy. In Perea (2012) it is shown, moreover, that the same strong correct beliefs assumption can be used to epistemically characterize perfect equilibrium and proper equilibrium in two-player static games,[2] whereas Perea and Predtetchinski (2016) show that it can be used to characterize subgame perfect equilibrium in two-player dynamic games as well. In that sense, the strong correct beliefs assumption can be viewed as a characterization of equilibrium reasoning in two-player games, and we use it as such in the second part of this paper. Moreover, we identify forward induction reasoning with common strong belief in rationality – the epistemic concept that characterizes extensive-form rationalizability.

The question thus becomes for which games the strong correct beliefs assumption is consistent with the conditions in common strong belief in rationality. In Theorem 5.3 we characterize that class of games by making use of extensive-form best response sets with unique beliefs – a refinement of the concept of extensive-form best response sets by Battigalli and Friedenberg (2012). We argue that only very few games meet the conditions in that characterization, and hence there are only few games where the strong correct beliefs assumption is compatible with common strong belief in rationality.

We next focus on games with perfect information and without relevant ties. On the basis of Theorem 5.3 we provide, in Theorem 6.1, a necessary condition for the consistency between the strong correct beliefs assumption and common strong belief in rationality, which is easy to verify. We prove that for every two-player dynamic game with perfect information and without relevant ties, the backward induction path must necessarily reach all histories that are consistent with both players' rationality if the strong correct beliefs assumption is to be consistent with common strong belief in rationality. As this condition will only very rarely be met, it follows that there are only very few games with perfect information where the strong correct beliefs assumption is consistent with common strong belief in rationality.

[2] Appears as an exercise in Perea (2012).

Despite the inconsistency between the correct beliefs assumption and common strong belief in rationality, there are forward induction *equilibrium* concepts in the literature where the correct beliefs assumption 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, 1992a), *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 8 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 8 that none of the forward induction equilibrium concepts above is able to uniquely select the intuitive forward induction strategy (d, g) for player 2 in the game of Fig. 1. The reason is that, if player 2 at h_1 wishes to rationalize the "surprising" move b 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 (d, g) for player 2.

Such exogenous restrictions on 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.

Both in terms of conclusions and methodology, our work is related to Reny (1992b, 1993) who shows that there are only very few games where "common belief in rationality" is possible at all "relevant histories". Here, the relevant histories are those that are consistent with both players' rationality, and where there is no dominant choice. Reny (1993) formalizes the idea of "common belief in rationality at a collection of histories" by the notion of a *jointly rational belief system,* which turns out to be rather similar to our concept of extensive-form best response sets with unique beliefs. We show in Theorem 7.2 that if a collection of histories is reached by an extensive-form best response set with unique beliefs, then this collection of histories can also be sustained by a jointly rational belief system. By combining this result with Theorem 5.3, it follows that for every game where the strong correct beliefs assumption is consistent with common strong belief in rationality, there is a jointly rational belief system for all relevant histories. That is, "common belief in rationality" will be possible at all relevant histories in this case.

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 an epistemic model with types. We use this epistemic model to formalize the correct beliefs assumption and the concept of common strong belief in rationality. In Section 4 we show that in some dynamic games the correct beliefs assumption is inconsistent with the first two layers of common strong belief in rationality. In Section 5 we formalize the strong correct beliefs assumption, characterize the class of two-player dynamic games for which the strong correct beliefs assumption is consistent with common strong belief in rationality, and show that this class is actually very small. In Section 6 we concentrate on the class of perfect information games, derive an easy necessary condition that must be met if the strong correct beliefs assumption is consistent with common strong belief in rationality, and argue that this condition is only very rarely satisfied. In Section 7 we discuss the relation between our approach and Reny's notion of jointly rational belief systems. In Section 8 we discuss some forward induction *equilibrium* concepts that have been proposed in the literature, and explain why these concepts, in the game of Fig. 1, fail to uniquely select the intuitive forward induction strategy (d, g) for player 2. Section 9 contains the proofs.

2. Dynamic games

In this paper we will restrict 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.

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'. 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, represented 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.

In this paper we interpret *strategies* as *plans of action* (Rubinstein, 1991). That is, choices for player i are only prescribed at those non-terminal histories $h \in H_i$ that are still *reachable*, given the choices prescribed at earlier histories. Formally, let $\hat{H}_i \subseteq H_i$ be a subcollection of histories where player i is active, and let s_i be a mapping that assigns to every history $h \in \hat{H}_i$ some available choice $s_i(h) \in C_i(h)$. Say that a history $h \in H_i$ is *reachable* under s_i if at all $h' \in \hat{H}_i$ preceding h, the selected choice $s_i(h')$ is the unique choice in $C_i(h')$ that leads to h. Finally, the mapping s_i, which assigns to every $h \in \hat{H}_i$ some available choice, is called a *strategy* if $\hat{H}_i$ contains exactly those histories in H_i that are reachable under s_i.

For a given non-terminal history h and a player i, let $S_i(h)$ denote the set of strategies for player i under which h is reachable. That is, $s_i \in S_i(h)$ if and only if at every $h' \in H_i$ preceding h, the strategy s_i prescribes the unique choice in $C_i(h')$ that leads to h.

3. Common strong belief in rationality

In this section we give a formal definition of the correct beliefs assumption and the forward induction concept of common strong belief in rationality. Before doing so, we first show how we can efficiently encode belief hierarchies by means of epistemic models with types.

3.1. Epistemic model

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 the opponent's 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 finite two-player dynamic game G with observable past choices. 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 $S_j(h)$ is the set of j's strategies under which h is reachable. 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. By $\beta_i^M(t_i)$ we denote the conditional belief hierarchy induced by type t_i in the epistemic model M. See Battigalli and Siniscalchi (1999) for the precise definition of the induced belief hierarchy $\beta_i^M(t_i)$.

In order to not miss out on any belief hierarchies, we must make sure that the epistemic model at hand contains *all* belief hierarchies that we are interested in. This leads to the notion of a *terminal* epistemic model (cf. Friedenberg, 2010).

Definition 3.2 *(Terminal epistemic model).* Consider a finite two-player dynamic game G with observable past choices, and an epistemic model $M = (T_i, b_i)_{i \in I}$ for G. The epistemic model M is terminal if for every other epistemic model $\hat{M} = (\hat{T}_i, \hat{b}_i)_{i \in I}$ for G, every player i, and every type $\hat{t}_i \in \hat{T}_i$, there is some type $t_i \in T_i$ with $\beta_i^M(t_i) = \beta_i^{\hat{M}}(\hat{t}_i)$.

Remember that $\beta_i^M(t_i)$ is the conditional belief hierarchy induced by type t_i in the epistemic model M, and similarly for $\beta_i^{\hat{M}}(\hat{t}_i)$. Hence, the condition above states that for every belief hierarchy that is induced by any type in any alternative epistemic model $\hat{M}$, there is already a type in M that induces exactly the same belief hierarchy. In other words, all possible belief hierarchies are already contained in M.

Battigalli and Siniscalchi (1999) have shown that for every finite dynamic game, we can always construct a terminal epistemic model which assumes (common belief in) Bayesian updating.[3] A similar construction can be employed to build a terminal epistemic model without Bayesian updating, as we use here.

[3] In fact, Battigalli and Siniscalchi (1999) construct for every finite dynamic game a *universal* epistemic model, which – in particular – is terminal.

3.2. Correct beliefs assumption

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 opponent is correct about the first-order belief he holds.

To formalize this condition within an epistemic model $M = (T_i, b_i)_{i \in I}$, let $b_i^1(t_i, h) \in \Delta(S_j(h))$ be the induced first-order belief for type t_i at history $h \in H_i$, and let $b_i^1(t_i) := (b_i^1(t_i, h))_{h \in H_i}$ be the induced collection of first-order beliefs. By

$$T_i[t_i] := \{t_i' \in T_i \mid b_i^1(t_i') = b_i^1(t_i)\}$$

we denote the set of types that share the same first-order beliefs as t_i, whereas

$$T_j(T_i[t_i]) := \{t_j \in T_j \mid b_j(t_j, h)(S_i(h) \times T_i[t_i]) = 1 \text{ for all } h \in H_j\}$$

is the set of types for player j that believe, throughout the game, that player i's first order belief is $b_i^1(t_i)$.

Definition 3.3 *(Correct beliefs assumption).* Consider a finite two-player dynamic game G with observable past choices, and an epistemic model $M = (T_i, b_i)_{i \in I}$ for G. Type $t_i \in T_i$ satisfies the correct beliefs assumption if $b_i(t_i, h)(S_j(h) \times T_j(T_i[t_i])) = 1$ for all histories $h \in H_i$.

That is, throughout the game type t_i assigns probability 1 to the event that player j, throughout the game, assigns probability 1 to his actual first-order belief.

Note that our correct beliefs assumption is somewhat different from the correct beliefs assumption that Aumann and Brandenburger (1995) use to epistemically characterize Nash equilibrium in two-player games. Aumann and Brandenburger's condition states that both players must be correct about the opponent's *actual* first-order beliefs. By doing so, they consider a state of the world where the actual first-order beliefs of both players are given. They thus take the viewpoint of the *analyst* who knows both players' first-order beliefs at that state.

In contrast, we take a *one-player perspective* in this paper, and not the perspective of an analyst. That is, we describe all beliefs from the viewpoint of a single player, say player i, who does not know the actual beliefs held by player j. Because of this, we cannot speak of the "actual first-order beliefs held by player j", and can therefore not formalize statements such as "player i is correct about j's first-order beliefs". What we *can* formalize is the statement that "player i believes that player j is correct about i's first-order beliefs", since player i knows his own first-order beliefs by positive introspection. This is precisely the approach we take here. Consequently, our correct beliefs assumption involves one order of belief more than Aumann and Brandenburger's condition, because we view everything from the perspective of a single player, and not from the perspective of an omniscient analyst.

3.3. Common strong belief in rationality

The epistemic concept 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 he 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 $s_i \in S_i(h)$. 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 pairs. 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 3.4 *(Rational choice).* Consider a type t_i for player i, a history $h \in H_i$ and a strategy $s_i \in S_i(h)$. 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 $s_i' \in S_i(h)$. Strategy s_i is rational for type t_i if it is so at every history $h \in H_i$ that is reachable under s_i.

We next define the notion of strong belief.

Definition 3.5 *(Strong belief).* Consider a type t_i within a terminal 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 epistemic concept of common strong belief in rationality can now be defined as follows.

Definition 3.6 *(Common strong belief in rationality).* Consider a finite two-player dynamic game G with observable past choices, and a terminal epistemic model $M = (T_i, b_i)_{i \in I}$ for G. 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 $R_i^0 := \{(s_i, t_i) \in S_i \times T_i^0 \mid s_i \text{ rational for } t_i\}$.

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,

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

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

For every $k \geq 1$, we say that a type t_i expresses up to k-fold strong belief in rationality if $t_i \in T_i^k$. We say that a type t_i expresses *common strong belief in rationality* if $t_i \in T_i^\infty$. Our definition of common strong belief in rationality is almost identical to the definition put forward by Battigalli and Siniscalchi (2002), except for the fact that Battigalli and Siniscalchi additionally require the types to satisfy Bayesian updating. Shimoji and Watson (1998) show, however, that this difference does not matter for the *strategies* that are rational for types in T_i^∞. Obviously, the difference does matter for the set of *belief hierarchies* selected by the concept, which in our case is encoded by the set of types T_i^∞.

Similarly to Battigalli and Siniscalchi (2002), it can be shown 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). In view of the aforementioned result by Shimoji and Watson (1998), this epistemic characterization of extensive-form rationalizability holds independently of whether we impose Bayesian updating or not.

4. Inconsistency theorem

We show, by means of the game G in Fig. 1, that the correct beliefs assumption may be inconsistent with the first two layers of common strong belief in rationality. To that purpose, take an arbitrary terminal epistemic model $M = (T_i, b_i)_{i \in I}$ for G. We show that there is no type $t_1 \in T_1^2$, expressing up to 2-fold strong belief in rationality, that satisfies the correct beliefs assumption.

Take an arbitrary type $t_1^* \in T_1^2$. Then, 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, it follows that $R_2^1 \subseteq \{c, (d, g)\} \times T_2$. Hence,

$$b_1(t_1^*, \emptyset)(\{c, (d, g)\} \times T_2) = 1. \tag{1}$$

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 terminal, there is a type $t_1 \in T_1$ with $b_1(t_1, \emptyset)(\{(d, h)\} \times T_2) = b_1(t_1, h_2)(\{(d, h)\} \times T_2) = 1$. Since $((b, f), t_1) \in (S_1(h_1) \times T_1) \cap R_1^0$, 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 strategy (d, h) by player 2, there is no strategy $s_1 \in S_1(h_1)$ and no type $t_1 \in T_1[t_1^*]$, sharing the same first-order beliefs as t_1^*, for which $(s_1, t_1) \in R_1^0$. To see this, note that $S_1(h_1) = \{(b, e), (b, f)\}$, and both (b, e) and (b, f) yield t_1 an expected utility less than 3 – the utility it can guarantee at $\emptyset$ by choosing a. As $b_2(t_2, h_1)((S_1(h_1) \times T_1) \cap R_1^0) = 1$, this implies that $b_2(t_2, h_1)(S_1 \times T_1[t_1^*]) = 0$. Hence, we see that

$$b_2(t_2, h_1)(S_1 \times T_1[t_1^*]) = 0 \text{ for every type } t_2 \in T_2^1. \tag{2}$$

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 the set of types $T_1[t_1^*]$. This means, however, that type t_1^* cannot satisfy the correct beliefs assumption.

Since this holds for every type $t_1^* \in T_1^2$, we conclude that there is no type $t_1^* \in T_1^2$ that satisfies the correct beliefs assumption. We therefore obtain the following result.

Theorem 4.1 *(Inconsistency theorem). Consider the game G from Fig. 1. Then, for every terminal epistemic model $M = (T_i, b_i)_{i \in I}$ for G there is no type $t_1 \in T_1$ for player 1 that satisfies the correct beliefs assumption and expresses up to 2-fold strong belief in rationality.*

Hence, in the game of Fig. 1, equilibrium reasoning is incompatible with forward induction reasoning as embodied by the first two layers in common strong belief in rationality.

This theorem raises the natural question whether the correct beliefs assumption can already be in conflict with the *first* layer of common strong belief in rationality, that is, with the event that a player strongly believes in the opponent's rationality. The answer is *no*. For every game we can, for both players i, always find a type t_i that (a) always believes that j always assigns probability 1 to his *actual* type t_i, and (b) strongly believes in j's rationality. Indeed, consider any type t_i' that strongly believes in j's rationality. We can then transform t_i' into a new type t_i that coincides with t_i in its first- and second-order conditional beliefs, but possibly differs in its third- and higher order beliefs by imposing that t_i should only consider types for player j that assign probability 1 to his actual type t_i. Since t_i holds the same first- and second-order beliefs as t_i', and t_i' strongly believes in j's rationality, type t_i strongly believes in j's rationality as well. Moreover, t_i satisfies, by construction, the correct beliefs assumption. Therefore, the conflict between the correct beliefs assumption and common strong belief in rationality can only occur at the second layer of common strong belief in rationality, or higher.

The theorem above immediately raises the following question: Can we characterize those games for which equilibrium reasoning is compatible with forward induction reasoning? That will be the purpose of the following section.

5. When correct beliefs are consistent with forward induction

In this section we will ask for which games equilibrium reasoning *is* consistent with forward induction reasoning. As explained in the introduction, we identify equilibrium reasoning with a *strong correct beliefs assumption,* that we will formally define below, and we identify forward induction reasoning with the concept of common strong belief in rationality. The question thus becomes for which games the strong correct beliefs assumption is consistent with common strong belief in rationality. We will characterize this class of games in Theorem 5.3, and argue that this class contains only very few games. 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).

In this section we start by formally defining the strong correct beliefs assumption, after which we introduce the new notion of extensive-form best response sets with unique beliefs. Finally, we use the latter to characterize the class of games for which the strong correct beliefs assumption is compatible with common strong belief in rationality.

5.1. Strong correct beliefs assumption

The strong correct beliefs assumption, which characterizes equilibrium reasoning in two-player games, states that player i believes that opponent j is correct about his entire belief hierarchy, and that player i believes that j believes that i is correct about j's entire belief hierarchy. To formalize this condition, recall that a type in an epistemic model is a way to encode a belief hierarchy. If we assume that different types in the epistemic model induce different belief hierarchies, then the strong correct beliefs assumption states that a type for player i believes that j is correct about his type, and that this type believes that j believes that i is correct about j's type.

For a given type $t_i \in T_i$, let $T_j(t_i)$ be set of types $t_j \in T_j$ for which $b_j(t_j, h)(S_i \times \{t_i\}) = 1$ at all $h \in H_j$. That is, $T_j(t_i)$ contains those types for player j that always assign full probability to i's type t_i. Say that type t_i *believes that j is correct about his type* if $b_i(t_i, h)(S_j \times T_j(t_i)) = 1$ for every $h \in H_i$. Let T_j^* be the set of types for j that believe that i is correct about j's type. Similarly, we say that type t_i *believes that j believes that i is correct about j's type* if $b_i(t_i, h)(S_j \times T_j^*) = 1$ for all $h \in H_i$.

Definition 5.1 *(Strong correct beliefs assumption).* Consider a finite two-player dynamic game G with observable past choices, and an epistemic model $M = (T_i, b_i)_{i \in I}$ for G.

We say that type t_i satisfies the strong correct beliefs assumption if t_i believes that j is correct about his type, and if t_i believes that j believes that i is correct about j's type.

In this definition we are thus implicitly assuming that the epistemic model is non-redundant, in the sense that different types induce different belief hierarchies.

5.2. 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 make h reachable. 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 is reachable under s_i.

Definition 5.2 *(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.

5.3. Characterization result

We will now characterize the class of dynamic games for which the strong correct beliefs assumption is consistent with common strong belief in rationality. Before we state our result, we need a few more definitions. We say that in a given game G the strong correct beliefs assumption is *consistent* with common strong belief in rationality, if there is a terminal epistemic model $M = (T_i, b_i)_{i \in I}$ for G, and for every player i a type $t_i \in T_i$ that expresses the strong correct beliefs assumption and common strong belief in rationality. 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 under which h is reachable.

Theorem 5.3 *(Consistency theorem). Consider a finite two-player dynamic game G with observable past choices. Then, the strong correct beliefs assumption is consistent with common strong belief in rationality in 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 $s_j \in D_j$ under which h is reachable.*

The proof of this theorem can be found in Section 9. This theorem implies that in "most" two-player dynamic games with observable past choices, the strong correct beliefs assumption is *inconsistent* with common strong belief in rationality. Indeed, assume that in the game G the strong correct beliefs assumption is consistent with common strong belief in rationality. Then, according to Theorem 5.3, 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 $s_j \in D_j$ under which h is reachable. 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 $s_j \in D_j$ under which h is reachable. 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", the strong correct beliefs assumption will be *inconsistent* with common strong belief in rationality.

On the other hand, the theorem shows that if the strong correct beliefs assumption is consistent with the *first two* layers of common strong belief in rationality, then it will also be consistent with *all* layers of common strong belief in rationality. To see this, consider an extensive-form best response set $D_1 \times D_2$ with unique beliefs, induced by conditional belief vectors b_1 and b_2. This describes a situation in which player i has the conditional belief vector b_i, believes that j has the conditional belief vector b_j, believes that j believes that i holds b_i, believes that j believes that i believes that j holds b_j, and so on. That is, player i holds a belief hierarchy that is entirely generated by the conditional belief vectors b_i and b_j, and therefore satisfies the strong correct beliefs assumption. Moreover, conditions (a), (b) and (c) in the definition of an extensive-form best response set with unique beliefs state that player i expresses up to two-fold strong belief in rationality. Overall, we thus see that an extensive-form best response set with unique beliefs can be interpreted as a situation where a player satisfies the strong correct beliefs assumption, and expresses up to two-fold strong belief in rationality.

Consequently, the latter statement in Theorem 5.3 describes situations where the strong correct beliefs assumption is consistent with the first two layers of common strong belief in rationality. Hence, Theorem 5.3 implies that, whenever the strong correct beliefs assumption is consistent with the first two layers of common strong belief in rationality, then it must be consistent with *all* layers of common strong belief in rationality.

In contrast, it can be shown that the strong correct beliefs assumption is always consistent with the backward induction concept of *common belief in future rationality* (Perea, 2014, see also Baltag et al., 2009 and Penta, 2015 for related concepts). Indeed, Perea and Predtetchinski (2016) show that in every finite two-player dynamic game with observable past choices, subgame perfect equilibrium can be epistemically characterized by common belief in future rationality, mutual belief in Bayesian updating, and the strong correct beliefs assumption. In particular, in all such games the strong correct beliefs assumption is consistent with common belief in future rationality. The intuitive reason is that in a backward induction concept there is no need to revise your belief about the opponent's belief hierarchy after a surprising move, as past moves need not be rationalized.

6. Games with perfect information

We have seen in the previous section that the strong correct beliefs assumption is very rarely consistent with common strong belief in rationality. In fact, we characterized the class of games for which the strong correct beliefs assumption *is* consistent with common strong belief in rationality – by relying on the concept of extensive-form best response sets with unique beliefs – and argued that this class is very small.

In this section we will focus on the special, yet important, class of dynamic games with *perfect information,* and show that for this class of games the strong correct beliefs assumption can only be consistent with common strong belief in rationality if the backward induction path reaches all histories that are consistent with both players' rationality. Since, in general, the set of histories that are consistent with both players' rationality is very large, this condition will be very rarely met. And hence we may conclude, on the basis of this result, that also within the class of games with perfect information, there are only very few games where the strong correct beliefs assumption is consistent with common strong belief in rationality.

The reason why we include this result in this paper is that the above condition – stating that the backward induction path reaches all histories that are consistent with both players' rationality – is very intuitive and easy to check, and hence this result is important both conceptually and practically.

In order to formally state this result, we need a few definitions. Consider a finite two-player dynamic game G with observable past choices. We say that G is with *perfect information* if only one player is active at the time. That is, for every non-terminal history h, the set of active players $I(h)$ contains only one player. The game G is said to be *without relevant ties* (Battigalli, 1997) if for every player i, every non-terminal history $h \in H_i$ where i is active, every two different choices $c_i, c_i' \in C_i(h)$, every terminal history z following c_i and every terminal history z' following c_i', we have that $u_i(z) \neq u_i(z')$. It is easy to see that in the absence of relevant ties the

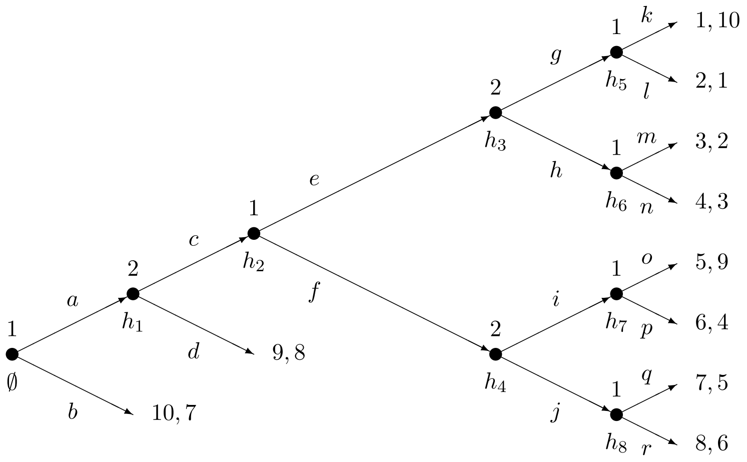


Fig. 2. Converse of Theorem 6.1 is false.

backward induction procedure selects a unique strategy for both players, and hence induces a unique backward induction path. Finally, recall that we called a strategy $s_i \in S_i$ *rational* if it is rational for at least one conditional belief vector. We say that a non-terminal history h is *consistent with both players' rationality* if h is reached by a strategy pair $(s_1, s_2) \in S_1 \times S_2$ where s_1 and s_2 are rational.

Theorem 6.1 *(Games with perfect information). Let G be a finite two-player dynamic game with perfect information and without relevant ties. If the strong correct beliefs assumption is consistent with common strong belief in rationality in G, then the unique backward induction path in G must reach all non-terminal histories that are consistent with both players' rationality.*

The proof, which can be found in Section 9, relies heavily on Theorem 5.3, Proposition 6 in Battigalli and Siniscalchi (2002) which shows that common strong belief in rationality leads to the extensive-form rationalizable strategies, and Theorem 4 in Battigalli (1997) which proves that in all perfect information games without relevant ties, every combination of extensive-form rationalizable strategies induces the backward induction path. Indeed, the main steps in the proof may be summarized as follows: If the strong correct beliefs assumption is consistent with common strong belief in rationality, then Theorem 5.3 guarantees that there is an extensive-form best response set with unique beliefs $D_1 \times D_2$ which reaches all histories that are consistent with both players' rationality. It can be shown that all strategies in D_1 and D_2 are possible under common strong belief in rationality, and hence Proposition 6 in Battigalli and Siniscalchi (2002) implies that all strategies in D_1 and D_2 are extensive-form rationalizable. But then, by Theorem 4 in Battigalli (1997), every strategy combination in $D_1 \times D_2$ induces the backward induction path. Since $D_1 \times D_2$ reaches all histories that are consistent with both players' rationality, we conclude that the backward induction path must reach all histories that are consistent with both players' rationality.

We will now show, by means of an example, that the other direction in Theorem 6.1 is not true. Consider the game in Fig. 2. It can easily be checked that this perfect information game is without relevant ties. Clearly, the backward induction path is b, leading to the terminal history with utilities 10 and 7. Moreover, as b is the only rational strategy for player 1, the only non-terminal history that is consistent with both players' rationality is $\emptyset$. Consequently, the backward induction path reaches all non-terminal histories that are consistent with both players' rationality.

Despite this fact, we will show that the strong correct beliefs assumption is not consistent with common strong belief in rationality in this game. In order to show this, we rely on Theorem 5.3. More precisely, we will prove that there is *no* extensive-form best response set $D_1 \times D_2$ with unique beliefs such that every history $h \in H_1$ that is consistent with 2's rationality is reachable under strategies in D_2.

Note first that (c, g, j) and (c, h, i) are rational strategies for player 2 in this game. Indeed, strategy (c, g, j) is rational for the conditional belief vector $b_2^{(c,g,j)}$ where

$$b_2^{(c,g,j)}(h_1) = (a, e, k, m), \quad b_2^{(c,g,j)}(h_3) = (a, e, k, m), \quad b_2^{(c,g,j)}(h_4) = (a, f, p, r).$$

Here, $b_2^{(c,g,j)}(h_1) = (a, e, k, m)$ denotes the conditional belief at h_1 that assigns probability 1 to 1's strategy (a, e, k, m), and similarly for the other two conditional beliefs. Similarly, strategy (c, h, i) is rational for the conditional belief vector $b_2^{(c,h,i)}$ where

$$b_2^{(c,h,i)}(h_1) = (a, f, o, q), \quad b_2^{(c,h,i)}(h_3) = (a, e, l, n), \quad b_2^{(c,h,i)}(h_4) = (a, f, o, q).$$

Since (c, g, j) and (c, h, i) are rational strategies for player 2, it follows that the non-terminal histories h_6 and h_8 in H_1 are both consistent with player 2's rationality.

We will show that there is *no* extensive-form best response set $D_1 \times D_2$ with unique beliefs such that both h_6 and h_8 are reachable under strategies in D_2. Consider an arbitrary extensive-form best response set $D_1 \times D_2$ with unique beliefs. Then, there is a conditional belief vector b_2^* for player 2 such that D_2 contains exactly those strategies for player 2 that are rational for b_2^*.

Now suppose, contrary to what we want to prove, that both h_6 and h_8 are reachable under strategies in D_2. Note that (c, h, i) and (c, h, j) are the only strategies for player 2 under which h_6 is reachable. Since strategy d is always better than (c, h, j) at h_1, it follows that (c, h, j) cannot be rational for b_2^*, and hence $(c, h, j) \notin D_2$. Since, by our assumption, h_6 is reachable under a strategy in D_2, it must be that $(c, h, i) \in D_2$. Similarly, note that (c, g, j) and (c, h, j) are the only strategies for player 2 under which h_8 is reachable. As, by our argument above, $(c, h, j) \notin D_2$, and, by our assumption, h_8 is reachable under a strategy in D_2, it must be that $(c, g, j) \in D_2$. We thus conclude that both (c, g, j) and (c, h, i) are in D_2. Hence, both (c, g, j) and (c, h, i) must be rational for the same conditional belief vector b_2^*.

Since (c, g, j) is rational for b_2^*, it must hold that

$$u_2((c, g, j), b_2^*(h_1)) \geq u_2(d, b_2^*(h_1)) = 8.$$

This implies that

$$b_2^*(h_1)(S_1(h_2, e)) \geq \frac{8}{10}, \tag{4}$$

where $S_1(h_2, e)$ denotes the set of strategies in $S_1(h_2)$ that select the choice e at h_2.

Similarly, since (c, h, i) is rational for b_2^*, it must hold that

$$u_2((c, h, i), b_2^*(h_1)) \geq u_2(d, b_2^*(h_1)) = 8.$$

This implies that

$$b_2^*(h_1)(S_1(h_2, f)) \geq \frac{8}{9}, \tag{5}$$

where $S_1(h_2, f)$ denotes the set of strategies in $S_1(h_2)$ that select the choice f at h_2. Clearly, the conditions (4) and (5) are incompatible, and hence there is no extensive-form best response set $D_1 \times D_2$ with unique beliefs such that both h_6 and h_8 are reachable under strategies in D_2. Since h_6 and h_8 are consistent with 2's rationality, this implies, by Theorem 5.3, that the strong correct beliefs assumption is inconsistent with common strong belief in rationality in the game of Fig. 2. However, the unique backward induction path reaches all non-terminal histories that are consistent with both players' rationality. Hence, the converse of Theorem 6.1 does not hold.

7. Jointly rational belief systems

In the last two sections we have shown that there are only very few games where the strong correct beliefs assumption is consistent with common strong belief in rationality. Similarly, Reny (1992b, 1993) proves that the class of games where "common belief in rationality" is possible at all "relevant histories" is very small. Here, a history is called "relevant" if it is consistent with both players' rationality, and if no player has a dominant choice there. Although the two questions asked above are very different, we will see that Reny's approach is similar to ours in terms of the concept being used.

More precisely, Reny (1993) proposes *jointly rational belief systems* as a way to formalize the idea of "common belief in rationality at a collection of histories", and we will see that it is rather similar to the notion of *extensive-form best response sets with unique beliefs* that we have used in Section 5. This relation is confirmed by Theorem 7.2 below, where we show that every collection of histories that is reached by an extensive-form best response set with unique beliefs, is possible under a jointly rational belief system. From this result and Theorem 5.3 it then follows that for every game where the strong correct beliefs assumption is consistent with common strong belief in rationality, common belief in rationality will be possible at all relevant histories.

7.1. Definition

Formally speaking, Reny (1993) only considers two-player games with *perfect information,* but his notion of jointly rational belief systems can easily be extended to games with observable past choices, as we will do here. Contrary to our model, Reny (1993) assumes that players hold a conditional belief at *every* non-terminal history in the game, also at those where they are not active. To make Reny's model fully compatible with ours, we assume in this section that both players are active at *all* non-terminal histories in the game. This can be assumed without loss of generality, as we can let a player choose from a singleton choice set at those histories where in reality he is not active.

Definition 7.1 *(Jointly rational belief system).* Let $\hat{H} \subseteq H$ be a collection of non-terminal histories. A jointly rational belief system for $\hat{H}$ is a non-empty set $D_1 \times D_2 \subseteq S_1 \times S_2$ of strategy pairs such that for both players i,

$$D_i = \{s_i \in S_i \mid s_i \text{ is rational for some conditional belief vector } b_i \\ \text{with } b_i(h)(D_j) = 1 \text{ for all } h \in \hat{H}\}.$$

Battigalli and Siniscalchi (1999) show that jointly rational belief systems can be epistemically characterized by *rationality and common certainty of the opponent's rationality given* $\hat{H}$, confirming that it indeed formalizes the idea of "common belief in rationality at a collection of histories". If we choose $\hat{H} = \{\emptyset\}$, that is, $\hat{H}$ contains only the beginning of the game, then the induced concept is *common certainty of rationality at the beginning of the game* as studied in Ben-Porath (1997).

Battigalli and Siniscalchi (1999) prove, moreover, that the largest jointly rational belief system for $\hat{H}$ can be obtained by the following inductive procedure: For both players i, let

$$R_i^{\hat{H},0} := \{s_i \in S_i \mid s_i \text{ is rational for some conditional belief vector } b_i\},$$

and for every round $k \geq 1$ and both players i, let

$$R_i^{\hat{H},k} = \{s_i \in S_i \mid s_i \text{ is rational for some conditional belief vector } b_i \\ \text{with } b_i(h)(R_j^{\hat{H},k-1}) = 1 \text{ for all } h \in \hat{H}\}.$$

Then, the limit set $R_1^{\hat{H}} \times R_2^{\hat{H}} := \cap_{k\geq 0}(R_1^{\hat{H},k} \times R_2^{\hat{H},k})$, provided it is non-empty, is the largest jointly rational belief system for $\hat{H}$. We will use this property in the proof of Theorem 7.2 below.

A jointly rational belief system need not always exist for every collection of non-terminal histories $\hat{H}$. Indeed, Reny (1993) shows that there are only very few games where a jointly rational belief system exists for the collection of all relevant histories. In these cases, the limit set $R_1^{\hat{H}} \times R_2^{\hat{H}}$ above will be empty. Following Reny (1993), we can say that "common belief in rationality is possible at a collection of histories $\hat{H}$" precisely when there is a jointly rational belief system for $\hat{H}$.

7.2. Connection to our work

To see the similarity with our notion of extensive-form best response sets with unique beliefs, note that a set $D_1 \times D_2 \subseteq S_1 \times S_2$ of strategy pairs is an extensive-form best response set with unique beliefs precisely when for both players i there is a conditional belief vector b_i with $b_i(h)(D_j) = 1$ for every h at which $S_j(h) \cap D_j \neq \emptyset$, such that

$$D_i = \{s_i \in S_i \mid s_i \text{ is rational for } b_i\}.$$

This follows immediately from Definition 5.2. This characterization seems rather similar to, and in some sense "more restrictive than", the notion of jointly rational belief systems. Indeed, in Theorem 7.2 below we show that whenever a collection $\hat{H}$ is reached by an extensive-form best response set with unique beliefs, then there will be a jointly rational belief system for $\hat{H}$.

In this theorem, we say that a collection of non-terminal histories $\hat{H}$ is *reached* by an extensive-form best response set with unique beliefs, if there is an extensive-form best response set $D_1 \times D_2$ with unique beliefs such that every $h \in \hat{H}$ is reached by some strategy pair in $D_1 \times D_2$.

Theorem 7.2 *(Connection with jointly rational belief systems). Consider a finite two-player dynamic game G with observable past choices, and a collection $\hat{H} \subseteq H$ of non-terminal histories. If $\hat{H}$ is reached by an extensive-form best response set with unique beliefs, then there is a jointly rational belief system for $\hat{H}$.*

The proof can be found in Section 9. In combination with Theorem 5.3 this result implies that, whenever in a given game the strong correct beliefs assumption is consistent with common strong belief in rationality, then there will be a jointly rational belief system for all relevant histories.

We can actually show a little more. Recall that a strategy s_i is called *rational* if it is rational for some conditional belief vector b_i. Let

$$H^{rat} := \{h \in H \mid \text{there are rational strategies } s_1, s_2 \text{ such that } (s_1, s_2) \text{ reach } h\}$$

be the collection of non-terminal histories that are consistent with both players' rationality. Then we can show that for every game where the strong correct beliefs assumption is consistent with common strong belief in rationality, there will be a jointly rational belief system for H^{rat}, which includes the set of all relevant histories as a subset.

Corollary 7.3 *(Jointly rational belief system for all relevant histories). Consider a finite two-player dynamic game G with observable past choices. If the strong correct beliefs assumption is consistent with common strong belief in rationality in G, then there is a jointly rational belief system for H^{rat}.*

The proof is immediate. If the strong correct beliefs assumption is consistent with common strong belief in rationality, then Theorem 5.3 guarantees that there is an extensive-form best response set $D_1 \times D_2$ with unique beliefs such that for both players i, and every history $h \in H$ that is consistent with i's rationality, there is a strategy $s_i \in D_i \cap S_i(h)$. In particular, every $h \in H^{rat}$ will be reached by some strategy pair $(s_1, s_2) \in D_1 \times D_2$, and hence H^{rat} is reached by an extensive-form best response set with unique beliefs. By Theorem 7.2 above it then follows that there is a jointly rational belief system for H^{rat}.

The converse of Corollary 7.3 is not true. To see this, consider the game in Fig. 2 where $H^{rat} = \{\emptyset\}$. It may be verified that $\{b\} \times \{d\}$ is a jointly rational belief system for H^{rat}. However, as we have seen in the previous section, the strong correct beliefs assumption is not consistent with common strong belief in rationality in this game.

Corollary 7.3 implies that if the strong correct beliefs assumption is consistent with common strong belief in rationality, then in particular there is a jointly rational belief system for the collection of all *relevant* histories. Reny (1993) and Battigalli and Siniscalchi (1999) have shown, in turn, that for games with perfect information and without relevant ties, there is a jointly rational belief system for the collection of all relevant histories, if and only if, the backward induction path reaches all relevant histories. By combining these two results, we conclude that if in a game with perfect information and without relevant ties, the strong correct beliefs assumption is consistent with common strong belief in rationality, then the backward induction path must reach all relevant histories. This is consistent with our Theorem 6.1 which states that in this case, the backward induction path must reach all histories that are consistent with both players' rationality (and not only the relevant histories).

8. Forward induction equilibrium concepts

8.1. Comparison in terms of strategies

In Section 4 we have seen that the correct beliefs assumption, which is implicitly assumed by all equilibrium concepts, is inconsistent with the first two layers of common strong belief in rationality. At the same time, the literature offers a broad spectrum of forward induction *equilibrium* concepts which – by construction – *are* consistent with the correct beliefs assumption, 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 strategy (d, g) of player 2 in the game of Fig. 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 Fig. 1. Why is (d, g) the intuitive forward induction strategy of player 2 here? If player 2 must make a choice at h_1 he knows that player 1 has chosen b, and not a, at the beginning. But choosing b can only be optimal for player 1 at the beginning if he subsequently chooses f, and believes that player 2, with high probability, will make the irrational choice h. Hence, player 2 must conclude that player 1 will subsequently choose f. As such, the only natural forward induction strategy for player 2 is to choose (d, g).

Note that, in order for player 2 to carry out this forward induction reasoning, he must consider the possibility that player 1 ascribes a high probability to the *irrational* strategy (d, h). 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 strategy (d, h), and therefore none of these equilibrium concepts is able to uniquely select the intuitive forward induction strategy (d, g) 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, 1992a), *outcomes satisfying forward induction* (Govindan and Wilson, 2009) and *Man's forward induction equilibrium* (Man, 2012).[4] Of these concepts, the former three are formulated as refinements of *sequential equilibrium* (Kreps and Wilson, 1982),[5] the fourth and fifth are defined as refinements of *weak sequential equilibrium* (Reny, 1992a),[6] whereas the last is a refinement of *normal-form perfect equilibrium* (Selten, 1975).

The game of Fig. 1 has a unique sequential equilibrium, in which player 2 chooses c. Consequently, *justifiable sequential equilibrium, Cho's forward induction equilibrium* and *stable sets of beliefs* – being refinements of sequential equilibrium – will uniquely select the strategy c for player 2, which is *not* the intuitive forward induction strategy, as we have seen.

In every *weak* sequential equilibrium of the game in Fig. 1, player 1 assigns probability 0 to player 2 choosing (d, h). 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 probability to strategy (d, h). As such, in these concepts player 2 cannot give a plausible explanation at h_1 for the event that player 1 has chosen b and not a. 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_2. In particular, both *explicable equilibrium* and *outcomes satisfying forward induction* allow for the strategies c and (d, g) by player 2, and therefore fail to single out the intuitive forward induction strategy (d, g) for player 2.

[4] 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 et al., 1993).

[5] Strictly speaking, Hillas' concept of *stable sets of beliefs* imposes restrictions on *sets* of sequential equilibria, rather than on *individual* sequential equilibria. However, these conditions can be translated into conditions on individual sequential equilibria.

[6] 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.

Since every normal-form perfect equilibrium induces a weak sequential equilibrium (see Reny, 1992a, Proposition 1), player 1 must also assign probability 0 to player 2's irrational strategy (d, h) 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 strategies c and (d, g) by player 2, and thus fails to uniquely select the intuitive forward induction strategy (d, g) for player 2.

We thus conclude that none of the forward induction equilibrium concepts above uniquely selects the intuitive forward induction strategy (d, g) of player 2. What prevents these forward induction equilibrium concepts from uniquely selecting the intuitive forward induction strategy (d, g) 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 strategy c by player 2, the forward induction arguments in the first three concepts have no bite in the game of Fig. 1, and still uniquely lead to the backward induction strategy c for player 2.

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

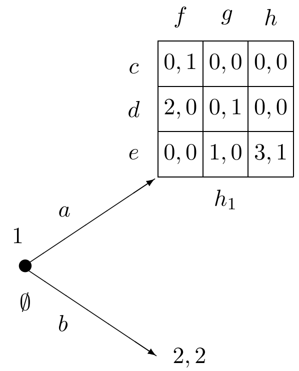
In contrast, the "pure" forward induction concept of *common strong belief in rationality* imposes no exogenous restrictions on the beliefs of the players, and therefore allows player 2 to reason about a scenario in which player 1 assigns a high probability to player 2's irrational strategy (d, h). This,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

8.2. *Comparison in terms of outcomes*

Some of the forward induction equilibrium concepts above, including Govindan and Wilson's (2009) notion of *outcomes satisfying forward induction,* focus on *outcomes* rather than strategies. For such concepts, the relevant question thus becomes whether it uniquely selects the "intuitive" forward induction *outcome(s),* rather than strategies. In the game of Fig. 1, for instance, Govindan and Wilson's concept still uniquely selects the intuitive forward induction outcome a, although it does not uniquely select player's 2's intuitive forward induction strategy (d, g).

However, we can also find games where common strong belief in rationality leads to a unique outcome which is not uniquely filtered by Govindan and Wilson's (2009) concept. Consider, for instance, the game in Fig. 3. Here, players 1 and 2 simultaneously choose from $\{c, d, e\}$ and $\{f, g, h\}$, respectively, at history h_1.

It can be shown that common strong belief in rationality uniquely leads to the outcome $((a, e), h)$, whereas Govindan and Wilson's concept allows for the outcome b. To see the former, note that player 2, at h_1, believes that player 1 will either choose d or e if he strongly believes in 1's rationality, as the strategy (a, c) will always yield player 1 less than 2. If player 2, in addition, chooses rationally at h_1, he will either choose g or h. Hence, if player 1 expresses



	f	g	h
c	$0,1$	$0,0$	$0,0$
d	$2,0$	$0,1$	$0,0$
e	$0,0$	$1,0$	$3,1$

Fig. 3. Govindan and Wilson's (2009) concept allows for outcomes that are not possible under common strong belief in rationality,

up to 2-fold strong belief in rationality, he must believe that player 2 will choose g or h. Player 1 should therefore choose e at h_1. If player 2 anticipates on this, he must choose h at h_1. But if player 1 expects player 2 to choose h, his optimal strategy is to choose (a, e). Hence, the only outcome to be expected under common strong belief in rationality is $((a, e), h)$.

However, we will show that the outcome b satisfies forward induction according to Govindan and Wilson's concept.[7] Consider the combination $(\sigma_1, \sigma_2, \beta_1, \beta_2)$ of behavioral strategies and beliefs, where

$$\sigma_1 = (b, d) \text{ and } \sigma_2 = g$$

are the players' behavioral strategies, and

$$\beta_1 = g \text{ and } \beta_2 = (a, d)$$

are the players' beliefs. Here, $\beta_1 = g$ represents the conditional belief vector for player 1 where he assigns probability 1 to player 2's strategy g at histories $\emptyset$ and h_1. Similarly, $\beta_2 = (a, d)$ is the conditional belief vector for player 2 where he assigns probability 1 to player 1's strategy (a, d) at h_1.

It may be verified that $(\sigma_1, \sigma_2, \beta_1, \beta_2)$ is a *weak sequential equilibrium*[8] leading to the outcome b. Moreover, player 1's strategy (a, d) is *relevant* for the outcome b, in Govindan and Wilson's terminology, since this strategy is optimal in another weak sequential equilibrium $(\sigma_1', \sigma_2', \beta_1', \beta_2')$, with

$$\sigma_1' = (b, c),\ \sigma_2' = f,\ \beta_1' = f,\ \beta_2' = (a, c),$$

that leads to the same outcome b. Therefore, player 2's belief β_2 in the first weak sequential equilibrium satisfies the forward induction condition in Govindan and Wilson, as it only assigns positive probability to strategies of player 1 that are relevant for the outcome b. As such, the outcome b satisfies forward induction in Govindan and Wilson's terminology.

[7] We refer the reader to Govindan and Wilson (2009) for the precise definitions.

[8] Note that player 1's behavioral strategy (b, d) is not optimal at h_1 given his belief β_1 there. This, however, is not required by weak sequential equilibrium, since the history h_1 is precluded by the behavioral strategy (b, d) itself. Weak sequential equilibrium requires a behavioral strategy only to be optimal at histories that it does not preclude from being reached.

We thus see that in the game of Fig. 3, Govindan and Wilson's (2009) forward induction concept allows an outcome, b, which is not possible under common strong belief in rationality. The main reason is that Govindan and Wilson's forward induction condition only invokes two reasoning steps: that player 2, at h_1, only assigns positive probability to relevant strategies of player 1, and that player 1 anticipates on this reasoning by player 2. Common strong belief in rationality, on the other hand, assumes more than two reasoning steps by the players in this game.

8.3. *Common strong belief in rationality with exogenous restrictions*

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 terminal.* That is, 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 non-terminal 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 Fig. 1, such an exogenous restriction could be that we only allow for types of player 1 that assign probability 0 to 2's irrational strategy (d, h). With such an exogenous restriction, *common strong belief in rationality* loses all of its bite in the game of Fig. 1, as player 2, at h_1, can no longer rationalize the event that player 1 has chosen b. Consequently, common strong belief in rationality would allow player 2 to choose either c or (d, g), and hence would no longer uniquely select the intuitive forward induction strategy (d, g) 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.

9. Proofs

Proof of Theorem 5.3. Let $M = (T_i, b_i)_{i\in I}$ be an arbitrary terminal 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 the strong correct beliefs assumption is consistent with common strong belief in rationality 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 $s_j \in D_j$ under which h is reachable.

As the strong correct beliefs assumption is consistent with common strong belief in rationality at G, there is for both players i a type $t_i^* \in T_i^\infty$ that satisfies the strong correct beliefs assumption. Consider a player i 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^* satisfies the strong correct beliefs assumption, type t_i^* believes that j is correct about his type. That is, $b_i(t_i^*, h)(S_j \times T_j(t_i^*)) = 1$ for every $h \in H_i$, where

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

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, for every $t_j \in T_j(t_i^*)$ there is some $h \in H_i$ with $b_i(t_i^*, h)(S_j \times \{t_j\}) < 1$. Take an arbitrary $t_j \in T_j(t_i^*)$. Then, we know from the above that there is some $h \in H_i$ with

$$b_i(t_i^*, h)(S_j \times \{t_j\}) < 1. \tag{6}$$

Moreover, as $t_j \in T_j(t_i^*)$ we know that

$$b_j(t_j, h')(S_i \times \{t_i^*\}) = 1 \text{ for all } h' \in H_j. \tag{7}$$

From (6) and (7) we conclude that type t_j does not believe that i is correct about j's type. 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 type. This, however, is a contradiction, since t_i^* satisfies the strong correct beliefs assumption. 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 $s_j \in D_j^*$ under which h is reachable.

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 $s_j \in D_j^*$ under which h is reachable. 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 $s_j \in D_j^*$ under which h is reachable.

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

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

which implies that there must be a strategy $s_j \in D_j^*$ under which h is reachable. 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 $s_j \in D_j^*$ under which h is reachable.

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 $s_j \in D_j^*$ under which h is reachable. 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 $s_j \in D_j$ under which h is reachable. We show that the strong correct beliefs assumption is consistent with common strong belief in rationality 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 $s_j \in D_j^*$ under which h is reachable. 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,

$$b_i(t_i^*, h)(\{(s_j, t_j^*)\}) := b_i^*(h)(s_j) \tag{8}$$

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 terminal epistemic model, such types t_1^* and t_2^* exist. By (8) 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^* satisfy the strong correct beliefs assumption.

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 $k \geq 0$, where T_i^k is defined as in Definition 3.6.

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 $s_j \in D_j^*$ under which h is reachable. 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 (8) it follows that

$$b_i(t_i^*, h)(D_j^* \times \{t_j^*\}) = 1. \tag{9}$$

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

$$D_j^* \times \{t_j^*\} \subseteq R_j^{k-1}. \tag{10}$$

If we combine (9) and (10), 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 satisfy the strong correct beliefs assumption. Hence, the strong correct beliefs assumption is consistent with common strong belief in rationality at G. This completes the proof of the theorem. □

Proof of Theorem 6.1. Suppose that the strong correct beliefs assumption is consistent with common strong belief in rationality in G. We must show that in G, every non-terminal history h that is consistent with both players' rationality lies on the backward induction path.

By Theorem 5.3, there is an extensive-form best response set $D_1 \times D_2$ with unique beliefs, such that for both players i, and every non-terminal history $h \in H_i$ that is consistent with j's rationality, there is a strategy $s_j \in D_j$ under which h is reachable. Let b_1^* and b_2^* be the conditional belief vectors associated with $D_1 \times D_2$. That is, for both players i we have that D_i is the set of strategies that are rational for b_i^*, and b_i^* strongly believes D_j.

Consider an arbitrary terminal epistemic model $M = (T_i, b_i)_{i \in I}$ for G. Let $t_1^* \in T_1$ and $t_2^* \in T_2$ be types such that for both players i,

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

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 terminal epistemic model, such types t_1^* and t_2^* exist in M.

Following the proof of Theorem 5.3, we know that the types t_1^* and t_2^* express common strong belief in rationality. Consider, for a given player i, a strategy $s_i \in D_i$. Then, by construction, s_i is rational for the conditional belief vector b_i^*. Since b_i^* is the first-order belief of type t_i^*, it follows that s_i is also rational for type t_i^*. Hence, we conclude that all strategies in D_i are rational for the type t_i^* which expresses common strong belief in rationality. By Proposition 6 in Battigalli and Siniscalchi (2002), it then follows that all strategies in D_i are extensive-form rationalizable in the sense of Pearce (1984) and Battigalli (1997).

Consider now an arbitrary strategy pair $(s_1, s_2) \in D_1 \times D_2$. Since both s_1 and s_2 are extensive-form rationalizable, (s_1, s_2) induces an extensive-form rationalizable outcome. Since the game G is a finite dynamic game with perfect information and without relevant ties, Theorem 4 in

Battigalli (1997) shows that the only extensive-form rationalizable outcome in G is the backward induction outcome. Hence, we conclude that every strategy pair $(s_1, s_2) \in D_1 \times D_2$ induces the unique backward induction path in G.

Take now an arbitrary non-terminal history h that is consistent with both players' rationality. For a given player i, consider the last history $h' \in H_i$ that weakly precedes h. Here, by "weakly precede" we mean that either h' precedes h, or $h' = h$. As h is consistent with both players' rationality, we know that h' is consistent with j's rationality. By our assumption above, we then know that there is a strategy $\hat{s}_j \in D_j$ under which h' is reachable. Hence, $\hat{s}_j \in D_j$ prescribes all the player j choices that precede h. As this holds for both players i, we conclude that there is a strategy pair $(s_1, s_2) \in D_1 \times D_2$ such that s_1 prescribes all the player 1 choices preceding h, and s_2 prescribes all the player 2 choices preceding h. But then, (s_1, s_2) leads to h. As $(s_1, s_2) \in D_1 \times D_2$, we know from above that (s_1, s_2) induces the backward induction path. Hence, h lies on the backward induction path.

We have thus shown that in G, every non-terminal history h that is consistent with both players' rationality lies on the backward induction path. This completes the proof. □

Proof of Theorem 7.2. Suppose that $\hat{H}$ is reached by an extensive-form best response set with unique beliefs. Then, there is an extensive-form best response set $D_1 \times D_2$ with unique beliefs such that every $h \in \hat{H}$ is reached by a strategy pair in $D_1 \times D_2$. By definition, for both players i there is a conditional belief vector b_i, with $b_i(h)(D_j) = 1$ for all h satisfying $S_j(h) \cap D_j \neq \emptyset$, such that $D_i = \{s_i \in S_i \mid s_i \text{ is rational for } b_i\}$. Since every $h \in \hat{H}$ is reached by a strategy pair in $D_1 \times D_2$, it follows in particular that $b_i(h)(D_j) = 1$ for all $h \in \hat{H}$.

For both players i, let $(R_i^{\hat{H},k})_{k \geq 0}$ be the sequence of strategy sets in Battigalli and Siniscalchi's (1999) inductive procedure yielding the limit set $R_1^{\hat{H}} \times R_2^{\hat{H}}$. We then know that $R_1^{\hat{H}} \times R_2^{\hat{H}}$, if non-empty, is the largest jointly rational belief system for $\hat{H}$. In order to show that there is a jointly rational belief system for $\hat{H}$, it is thus sufficient to show that $R_1^{\hat{H}} \times R_2^{\hat{H}}$ is non-empty. To that purpose we will show, by induction on k, that $D_1 \times D_2 \subseteq R_1^{\hat{H},k} \times R_2^{\hat{H},k}$ for all $k \geq 0$.

For $k = 0$, consider a player i and a strategy $s_i \in D_i$. Then, by construction, s_i is rational for the conditional belief vector b_i and hence $s_i \in R_i^{\hat{H},0}$. As this holds for both players i and all $s_i \in D_i$, it follows that $D_1 \times D_2 \subseteq R_1^{\hat{H},0} \times R_2^{\hat{H},0}$.

Now, take some $k \geq 1$ and assume that $D_1 \times D_2 \subseteq R_1^{\hat{H},k-1} \times R_2^{\hat{H},k-1}$. Take some player i and some $s_i \in D_i$. Then, s_i is rational for the conditional belief vector b_i which satisfies $b_i(h)(D_j) = 1$ for all $h \in \hat{H}$. By the induction assumption we know that $D_j \subseteq R_j^{\hat{H},k-1}$, and therefore $b_i(h)(R_j^{\hat{H},k-1}) = 1$ for all $h \in \hat{H}$. Hence, we conclude that $s_i \in R_i^{\hat{H},k}$. As this holds for both players i and every $s_i \in D_i$, it follows that $D_1 \times D_2 \subseteq R_1^{\hat{H},k} \times R_2^{\hat{H},k}$.

By induction, we thus conclude that $D_1 \times D_2 \subseteq R_1^{\hat{H}} \times R_2^{\hat{H}}$, and hence $R_1^{\hat{H}} \times R_2^{\hat{H}}$ is non-empty. It then follows that $R_1^{\hat{H}} \times R_2^{\hat{H}}$ is the largest jointly rational belief system for $\hat{H}$. In particular, there is a jointly rational belief system for $\hat{H}$, which completes the proof. □

References

Asheim, G.B., 2006. The Consistent Preferences Approach to Deductive Reasoning in Games. Theory and Decision Library. Springer, Dordrecht, The Netherlands.

Aumann, R., Brandenburger, A., 1995. Epistemic conditions for Nash equilibrium. Econometrica 63, 1161–1180.

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

Banks, J.S., Sobel, J., 1987. Equilibrium selection in signaling games. Econometrica 55, 647–661.

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

Battigalli, P., Friedenberg, A., 2012. Forward induction reasoning revisited. Theor. Econ. 7, 57–98.

Battigalli, P., Siniscalchi, M., 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. J. Econ. Theory 88, 188–230.

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

Ben-Porath, E., 1997. Rationality, Nash equilibrium and backwards induction in perfect information games. Rev. Econ. Stud. 64, 23–46.

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

Brandenburger, A., Dekel, E., 1987. Rationalizability and correlated equilibria. Econometrica 55, 1391–1402.

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

Cho, I.-K., 1987. A refinement of sequential equilibrium. Econometrica 55, 1367–1389.

Cho, I.-K., Kreps, D.M., 1987. Signaling games and stable equilibria. Q. J. Econ. 102, 179–221.

Friedenberg, A., 2010. When do type structures contain all hierarchies of beliefs? Games Econ. Behav. 68, 108–129.

Govindan, S., Wilson, R., 2009. On forward induction. Econometrica 77, 1–28.

Grossman, S.J., Perry, M., 1986. Perfect sequential equilibrium. J. Econ. Theory 39, 97–119.

Harsanyi, J.C., 1967–1968. Games with incomplete information played by "bayesian" players, I–III'. Manag. Sci. 14, 159–182, 320–334, 486–502.

Hillas, J., 1994. Sequential equilibria and stable sets of beliefs. J. Econ. Theory 64, 78–102.

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

Mailath, G.J., Okuno-Fujiwara, M., Postlewaite, A., 1993. Belief-based refinements in signalling games. J. Econ. Theory 60, 241–276.

Man, P.T.Y., 2012. Forward induction equilibrium. Games Econ. Behav. 75, 265–276.

McLennan, A., 1985. Justifiable beliefs in sequential equilibria. Econometrica 53, 889–904.

Myerson, R.B., 1978. Refinements of the Nash equilibrium concept. Int. J. Game Theory 7, 73–80.

Nash, J.F., 1950. Equilibrium points in N-person games. Proc. Natl. Acad. Sci. USA 36, 48–49.

Nash, J.F., 1951. Non-cooperative games. Ann. Math. 54, 286–295.

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

Penta, A., 2015. Robust dynamic implementation. J. Econ. Theory 160, 280–316.

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

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

Perea, A., 2014. Belief in the opponents' future rationality. Games Econ. Behav. 83, 231–254.

Perea, A., Predtetchinski, A., 2016. An Epistemic Approach to Stochastic Games. Working Paper.

Reny, P.J., 1992a. Backward induction, normal form perfection and explicable equilibria. Econometrica 60, 627–649.

Reny, P.J., 1992b. Rationality in extensive-form games. J. Econ. Perspect. 6, 103–118.

Reny, P.J., 1993. Common belief and the theory of games with perfect information. J. Econ. Theory 59, 257–274.

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

Selten, R., 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragezeit. Z. Gesammte Staatswiss. 121, 301–324, 667–689.

Selten, R., 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. Int. J. Game Theory 4, 25–55.

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

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