PLAUSIBILITY ORDERINGS IN DYNAMIC GAMES

Andrés Perea

Economics and Philosophy / Volume 30 / Issue 03 / November 2014, pp 331 - 364
DOI: 10.1017/S0266267114000285, Published online: 18 September 2014

Economics and Philosophy, **30** (2014) 331–364 © Cambridge University Press
doi:10.1017/S0266267114000285

PLAUSIBILITY ORDERINGS IN DYNAMIC GAMES

ANDRÉS PEREA
Maastricht University, Holland

In this paper we explore game-theoretic reasoning in dynamic games within the framework of *belief revision theory*. More precisely, we focus on the forward induction concept of 'common strong belief in rationality' (Battigalli and Siniscalchi (2002) and the backward induction concept of 'common belief in future rationality' (Baltag *et al*. 2009; Perea 2014). For both concepts we investigate whether the *entire* collection of selected belief revision policies for a player can be characterized by a *unique* plausibility ordering. We find that this is indeed possible for 'common strong belief in rationality', whereas this may be impossible in some games for 'common belief in future rationality'.

1. INTRODUCTION

Belief revision plays a fundamental role in human decision making, and determines to a large extent the choices we make. Indeed, the beliefs we hold today may be contradicted by new evidence tomorrow, and at that point in time we must be prepared to change our beliefs as to accommodate the new piece of information. The choices we make tomorrow will therefore crucially depend upon *how* we revise these beliefs.

So, in order to understand how people make decisions in a dynamic environment we must first investigate how they may revise their beliefs upon observing new facts. Of course there are many different ways of changing beliefs, and in order to develop a meaningful theory of belief

I would like to thank Alexandru Baltag, Giacomo Bonnano, Amanda Friedenberg, Hannes Leitgeb, Hans Rott, Sonja Smets and two anonymous referees for helpful comments.

revision we must present a list of desirable properties that a belief revision process should satisfy. In the 1980s, Alchourrón *et al.* (1985) presented one such list of properties – known today as the AGM-axioms – that would have a fundamental impact on the development of belief revision theory. The class of belief revision policies satisfying the AGM-axioms still serves as a central model of belief change in many works in various different areas.[1]

Some years later, Grove (1988) proved a beautiful characterization result which states that the belief revision policies satisfying the AGM-axioms are precisely those that can be derived from some *plausibility ordering* over states of the world. By the latter we mean that the decision-maker ranks all possible states of the world in terms of their (subjective) plausibility, and upon receiving a new piece of information restricts his new belief to the most plausible states that are compatible with the new information. This plausibility ordering not only determines the decision-maker's initial belief – namely the set of states he deems most plausible overall – but also how the decision-maker changes his belief in case the new information completely – or partially – contradicts his initial belief. In this case, the decision-maker restricts his attention to a smaller set of states – namely those that are not ruled out by the new information – and among these states he selects those that he finds most plausible. This will then serve as his new, revised belief. In my view, plausibility orderings provide a very natural way of inducing a belief revision policy, and Grove's representation theorem confirms that the AGM-axioms indeed establish an intuitive list of postulates, leading to a natural class of belief revision rules.

Belief revision is of special importance in *dynamic games*, where players may learn new facts about the past behaviour of their opponents during the game. In such cases, players may need to revise their beliefs about the opponents' strategy choices, and the eventual choices made by the players will crucially depend on *how* they revise these beliefs. As an illustration, consider the dynamic game in Figure 1.

For player 2 it seems reasonable to *initially* believe – before anything has happened – that player 1 will choose b and end the game immediately. To see this, note that for player 2 it is irrational to choose g. Hence, if player 1 believes that player 2 would choose rationally upon choosing a, then player 1 expects not to get more than 2 by choosing a, and therefore would rather choose b.

But what would player 2 *do* in this game? If it is player 2's turn to make a move, he knows that player 1 has chosen a, and not b, so player 2

[1] Despite their intuitive appeal, some authors have criticized the AGM-axioms for being too conservative. See, for instance, Levi (2013) and Kevin Kelly's work on simplicity and Occam's razor.

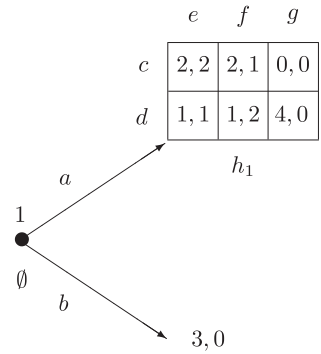

FIGURE 1. Belief revision matters for choices

has to revise his initial belief about player 1. But how? We will describe two plausible belief revision scenarios for player 2 in this game, leading to two different choices.

In the first scenario, player 2 believes that choosing a was a conscious, optimal choice for player 1. In that case, however, player 2 must believe that player 1 will subsequently choose d, as this is the only way for player 1 to obtain more than 3 – the utility he could have guaranteed by choosing b at the beginning. So, player 2 will respond by choosing f. Note that in this scenario, player 2, upon observing a, can no longer believe that player 1 believes that player 2 will choose rationally after a . Namely, in order to rationalize player 1's move a, player 2 must believe that player 1 ascribes a high probability to player 2 making the irrational choice g, as only then can player 1 achieve more than 3 by choosing a. This belief revision scenario corresponds to the forward induction concept of 'common strong belief in rationality' as developed by Battigalli and Siniscalchi (2002), and which is based upon the 'extensive-form rationalizability procedure' by Pearce (1984). The main idea in this concept is that a player, when he observes an unexpected move by his opponent, tries to interpret this move as being part of an optimal strategy, whenever this is possible. This is precisely what player 2 does in the game of Figure 1 under the belief revision scenario described above, when he observes the unexpected move a by player 1. He interprets a as being part of an optimal strategy by player 1, but then he must believe that player 1 will choose the follow-up action d, and hence player 2 will choose f himself.

This is not the only plausible way for player 2 to revise his belief, however. If he observes that player 1 has – surprisingly – chosen a, he could also believe that this was a mistake by player 1, but that player 1 will still choose rationally in the game that lies ahead, and that player 1 still believes that player 2 will choose rationally in the remainder of

the game. In that case, player 2 will believe that player 1 believes that player 2 will not choose g. Hence, player 2 will believe that player 1 chooses c after a – and not d as in the belief revision scenario above. As a consequence, player 2 will respond by choosing e – and not f as in the scenario above. This second belief revision scenario is implicit in the backward induction concept of 'common belief in future rationality' as proposed by Perea (2014) and Baltag *et al.* (2009). The key condition in this concept is that a player, upon observing an unexpected move by his opponent, always believes that the opponent will choose rationally from now on, and that the opponent believes that the other players will also choose rationally from now on, and so forth. However – and that is the main difference with 'common strong belief in rationality' – the player need not believe that the opponent's past choice was an optimal choice, even when believing so is possible. In fact, the player is free to believe that the unexpected move he observed was actually a mistake by the opponent. What is important is that the player believes that from now on everything is back to normal – that is, that the opponent will choose rationally from now on, and that the opponent believes that everybody else will choose rationally from now on, and so forth. In this sense the concept is entirely *forward looking*, as it only imposes conditions on how players reason about current and future moves, and not about past moves. That is why we call it a *backward induction* concept, as opposed to *forward induction* reasoning which requires players to also reason critically about opponents' past moves.

We thus see that the forward induction concept of 'common strong belief in rationality' and the backward induction concept of 'common belief in future rationality' do not only describe different belief revision scenarios for player 2 in the game above, but also lead to different *choices* for player 2. This shows that belief revision crucially matters for how players choose in a dynamic game.

Since belief revision is so important for the study of dynamic games, it seems only natural to embed the analysis of dynamic games within the framework of belief revision theory. But somewhat surprisingly, this approach has hardly been adopted so far in the game theory literature – some exceptions being the works by Bonanno (2009, 2011, 2013) and Baltag *et al.* (2009).

The purpose of this paper is to enhance this connection by building a bridge between the study of dynamic games on the one hand, and the idea of plausibility orderings in belief revision theory on the other hand. For this investigation we restrict our attention to the concepts of 'common strong belief in rationality' and 'common belief in future rationality' mentioned above. In general, these two concepts do not prescribe a *unique* belief revision policy for a player, but typically select for every dynamic game a *whole collection of belief revision policies* for this player. Of course,

these belief revision policies must share some common feature, since they all correspond to the same game-theoretic concept. The question we want to address for each of the two concepts is whether the entire collection of belief revision policies selected for a given player can be summarized by a *common plausibility ordering*. That is, can we find, for every player i, a *unique* plausibility ordering such that the belief revision policies selected for player i by the concept at hand are *precisely* those belief revision policies that respect this plausibility ordering. If that is true, then the common feature that these belief revision policies share is precisely this common plausibility ordering. In fact, the whole game-theoretic concept could then be summarized by *one* plausibility ordering for each of the players, which would constitute a very simple and natural representation of the concept. It would also reveal a clear intuition for the concept at hand, as the concept would require every player to revise his beliefs according to this common plausibility ordering – nothing more and nothing less.

We find that the collection of belief revision policies selected by 'common strong belief in rationality' can indeed by summarized by a single plausibility ordering for each of the players, whereas this is not always possible for 'common belief in future rationality' in some games. Moreover, we show in Theorem 6.4 what this plausibility ordering looks like for 'common strong belief in rationality'. In contrast, for the concept of 'common belief in future rationality' a unique plausibility ordering is often not enough to characterize *all* belief revision policies for a given player in the game. We provide an example for this in Section 7.2.

At the end of this paper we focus on the special class of games with *perfect information*, in which players move one at a time, and always observe precisely what their opponents have done so far. We show by means of a counterexample that even in such games, the collection of belief revision policies selected for a given player by 'common belief in future rationality' cannot be characterized by a unique plausibility ordering. However, we show that the concept can be refined to a stronger concept, 'common belief in rationality at future *and parallel* information sets', where these collections of belief revision policies *can* be characterized by a common plausibility ordering for this special class of games, provided there are no relevant ties in the game. Moreover, the latter concept, like 'common belief in future rationality', always induces the backward induction strategies in such games.

For the class of games with perfect information, Baltag *et al.* (2009) have defined the concept of 'common knowledge of stable belief in dynamic rationality', which has exactly the same spirit as 'common belief in future rationality', and show that it also uniquely yields the backward induction strategies in case there are no relevant ties. One difference with 'common belief in future rationality' is that their concept assumes, from

the beginning, that the belief revision policies are given by a unique plausibility ordering for every player. So, in that sense the concept is similar to 'common belief in rationality at future and parallel information sets' which is also characterized by a unique plausibility ordering for every player in perfect information games without relevant ties, and also uniquely leads to the backward induction strategies there. These insights show that for perfect information games, backward induction can be characterized by suitably chosen plausibility orderings.

The paper is organized as follows. In Section 2 we give a formal definition of a dynamic game and its associated strategies. In Section 3 we introduce hierarchies of conditional beliefs in dynamic games, and show how these can be encoded by means of an epistemic model with types. We also show how belief revision can be captured within this model. In Section 4 we give a definition of a 'reasoning context' for dynamic games, describing the possible belief hierarchies that a player can hold for any such game. The concepts of 'common strong belief in rationality' and 'common belief in future rationality' are thus special cases of a 'reasoning context'. In Section 5 we introduce plausibility orderings in dynamic games, and define what it means for a reasoning context to be characterized by a unique plausibility ordering for every player. This means that for every player, the whole collection of selected belief hierarchies can be summarized by a *unique* plausibility ordering for this player – precisely the idea we have discussed above. In Section 6 we formally introduce the concept of 'common strong belief in rationality' and show that it can always be characterized by a unique plausibility ordering for every player. In Section 7 we formally define the concept of 'common belief in future rationality' and demonstrate that it cannot always be characterized by a unique plausibility ordering for every player. In Section 8 we investigate the class of games with perfect information, as discussed above. In Section 9 we end with a discussion. All proofs are collected in the appendix.

2. DYNAMIC GAMES

2.1 A Model of Dynamic Games

In a *dynamic* game, players may have to choose more than once during the course of the game, and may partially or completely observe what other players have done in the past when it is their time to make a choice. Throughout this paper we assume 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. Formally, a *finite dynamic game G* consists of the following ingredients.

There is a finite set of players I. The instances where one or more players must make a choice are given by a finite set X 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 x where it is player i's turn to move. As player i may not fully observe what his opponents have done in the past, player i may not be able to distinguish x from other non-terminal histories. Formally, we model player i's information at x by an *information set* h that contains all non-terminal histories that, from player i's point of view, are indistinguishable from x. We denote by H_i the collection of all information sets for player i in the game. We assume that there is *perfect recall*, meaning that a player never forgets what he previously did, and what he previously knew about the opponents' past choices.

Consider a non-terminal history x at which player i must make a choice. By $C_i(x)$ we denote the finite set of choices that are available to player i at x. Let $h \in H_i$ be the information set for player i to which x belongs. As on the one hand, player i cannot distinguish x from other non-terminal histories in h, but on the other hand is assumed to *know* the set of choices available to him, we must require that $C_i(y) = C_i(x)$ for all non-terminal histories $y \in h$. But then, we may as well use the notation $C_i(h)$, specifying the (unique) set of choices available to player i at information set $h \in H_i$.

We explicitly allow for *simultaneous moves* in the dynamic game. That is, we allow for non-terminal histories at which several players make a choice. Formally, this means that for some non-terminal histories x there may be different players i and j, and information sets $h \in H_i$ and $h' \in H_j$, such that $x \in h$ and $x \in h'$. In this case, we say that the information sets h and h' are *simultaneous*. So, two information sets $h \in H_i$ and $h' \in H_j$ are simultaneous if they have a non-empty intersection. For instance, in the game of Figure 1 we see that players 1 and 2 simultaneously move at information set h_1. In that game, the information set for player 1 at that stage is identical to the information set for player 2 at that stage – both are equal to h_1. But in general there may also be *different* information sets $h \in H_i$ and $h' \in H_j$ that are simultaneous. Consider, for instance, two non-terminal histories x and y where both i and j make a choice. Suppose that player i knows at x that x has been reached. So, $h = \{x\}$ is an information set for player i. Suppose that player j does not know at x whether x or y has been reached. So, $h' = \{x, y\}$ is an information set for player j. Then, h and h' are simultaneous – yet different – information sets.

Consider a non-terminal history x where $I(x)$ is the set of active players. That is, $I(x)$ contains those players who must make a choice at x. Then, every combination of choices $(c_i)_{i \in I(x)}$ is assumed to move the game from the non-terminal history x to some other (terminal or

non-terminal) history y. These transitions can formally be described by a *move-function* m, which assigns to every non-terminal history x, and every combination of choices $(c_i)_{i \in I(x)}$, the (terminal or non-terminal) history $m(x)$ that follows.

We say that history y follows some other history x if y can be reached from x by a suitable sequence of choice combinations, given the move-function m. Similarly, we say that an information set h *follows* some other information set h' if there are histories $x \in h$ and $y \in h'$ such that x follows y. We say that information set h *weakly follows* h' if either h follows h', or h and h' are simultaneous. We assume, throughout this paper, that there is an *unambiguous ordering of the information sets* in the game. That is, if information set h follows information set h', then h' does not follow h. Or, equivalently, there cannot be histories $x, y \in h$, and histories $x', y' \in h'$ such that x follows x', and y' follows y.

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

2.2 Strategies

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

An argument that is often used in defence of the traditional definition of a strategy is that the choices specified at precluded information sets

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

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

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

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

Let us denote by S_i the set of all strategies for player i. Since the dynamic game G is finite, the set S_i will be finite as well. By $S := \times_{i \in I} S_i$ we denote the set of all strategy combinations, and for every player i we denote by $S_{-i} := \times_{j \in I \setminus \{i\}} S_j$ the set of strategy combinations for i's opponents. For a given information set $h \in H_i$, let $S(h)$ be the set of strategy combinations that reach h – that is, the set of strategy combinations $(s_j)_{j \in I}$ that reach some history in h if every player j carries out his strategy s_j. By $S_i(h)$ we denote the set of strategies s_i for player i for which there is some opponents' strategy combination $s_{-i} \in S_{-i}$ such that $(s_i, s_{-i}) \in S(h)$. We say that strategies in $S_i(h)$ *possibly reach* h. Similarly, $S_{-i}(h)$ denotes the set of strategy combinations $s_{-i} \in S_{-i}$ for which there is some strategy $s_i \in S_i$ such that $(s_i, s_{-i}) \in S(h)$. We say that strategy combinations in $S_{-i}(h)$ *possibly reach* h.

Consider some information set $h \in H_i$ for player i. As we assume that the game G has perfect recall, player i remembers at h each of his past choices, and hence h is preceded by a unique sequence of past choices for player i. So, $S_i(h)$ contains precisely those strategies that prescribe this unique sequence of player i choices preceding h. But then, it is not difficult to see that $S(h) = S_i(h) \times S_{-i}(h)$ for every $h \in H_i$.

3. MODELLING BELIEF HIERARCHIES

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

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

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

Consider, for instance, the game in Figure 1, and suppose that player 2 initially believes that player 1 chooses b. Then, if player 2 is required to make a choice at h_1, he knows that player 1 has chosen a, and hence his previous belief was wrong. Player 2 must therefore substitute it by a new conditional belief at h_1 that only considers strategies for player 1 that are still possible – namely (a, c) and (a, d).

Summarizing, we see that we need to model *conditional belief hierarchies* for a player, which specify at each of his information sets what he believes about the opponents' strategy choices, the opponents' first-order beliefs, the opponents' 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 information set $h \in H_i$ holds a belief about the opponents' strategies, the opponents' first-order beliefs, the opponents' second-order beliefs, and so on. In other words, this player holds at h a belief about the opponents' strategies and the opponents' conditional belief hierarchies. So, a conditional belief hierarchy for player i specifies at each of i's information sets a conditional belief about the opponents' strategy choices and the opponents' conditional belief hierarchies. If we substitute the word 'belief hierarchy' by the word 'type' – as Harsanyi did – then we obtain the following definition.

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

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

Recall that $S_{-i}(h)$ represents the set of opponents' strategy combinations that possibly reach h. By $T_{-i} := \times_{j \in I \setminus \{i\}} T_j$ we denote the set of opponents' type combinations. For every set X, we denote by $\Delta(X)$ the set of probability distributions on X. Clearly, player i must at h only assign positive probability to opponents' strategy combinations in $S_{-i}(h)$, as these are the only strategy combinations compatible with the event that h is reached. This explains the condition in (b) that $b_i(t_i, h) \in \Delta(S_{-i}(h) \times T_{-i})$. Note that in part (b) we require player i to hold a conditional belief also at $\emptyset$ – the beginning of the game – even when player i is not active there. Hence, we assume that every player holds an *initial belief* before the start of the game.

From now on, we will use the notation $H_i^* := H_i \cup \{\emptyset\}$. So, at every information set $h \in H_i^*$ type t_i holds a conditional probabilistic belief $b_i(t_i, h)$ about the opponents' strategies and types. In particular, type t_i holds conditional beliefs about the opponents' strategies. As every opponent's type holds conditional beliefs about the other players' strategies, every type t_i holds at every $h \in H_i$ also a conditional belief about the opponents' conditional beliefs about the other players' strategy choices. And so on. Since a type may hold different beliefs at different histories, a type may, during the game, revise his belief about the opponents' strategies, but also about the opponents' conditional beliefs. In fact, for a given type t_i within an epistemic model, we can *derive* the complete belief hierarchy it induces.

4. REASONING CONTEXTS

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

But if this is true, then we could attempt to formalize a reasoning context as follows: Take an arbitrary dynamic game G and an epistemic

model M. Then, a reasoning context selects for a given player a subset of types within M, representing those belief hierarchies that are 'allowed for' by the reasoning context. Although this may seem reasonable there is one major problem with this attempt, namely that the epistemic model at hand may not contain *all* belief hierarchies that we are interested in – some belief hierarchies that we would wish to select are simply not present in the epistemic model. In order to avoid this problem we assume the epistemic model to be *belief complete*[2](cf. Brandenburger 2003).

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

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

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

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

So, effectively, a reasoning context selects for every dynamic game a set of belief hierarchies for every player – those belief hierarchies that are deemed 'most plausible' by this reasoning context.

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

5. PLAUSIBILITY ORDERINGS

5.1 Plausibility Orderings in Dynamic Games

Plausibility orderings are a very natural way to generate, or characterize, belief revision policies by agents. Consider an agent whose space of uncertainty is given by a set X of possible states of the world. Now suppose that, before receiving any new information, this agent ranks the possible states in X according to a *plausibility ordering*. That is, for every pair of states x and y, the agent specifies which of these two states he deems more plausible, if any. If the agent subsequently receives new information revealing that the true state must be in $E \subseteq X$, then it makes intuitive sense for the agent to concentrate his conditional belief only on states in E that he deems 'most plausible'.

This approach plays an important role both in belief revision theory and counterfactual logic. Grove (1988) has shown that the belief revision policies that follow the AGM axioms are exactly those that can be characterized by plausibility orderings over states. So, in a sense, the AGM axioms for belief revision are logically equivalent to the use of plausibility orderings. In his paper, Grove uses *systems of spheres* instead of plausibility orderings, but we will see below that both approaches are equivalent. Lewis (1973) and Stalnaker (1968), on the other hand, use plausibility orderings to evaluate counterfactual statements. More precisely, they assume for every state x a plausibility ordering over states that deems x, and only x, as most plausible. According to the Lewis-Stalnaker theory, a conditional statement 'if p then q' is true at that state x if q is true at all 'most plausible p-states'. By the latter, we mean states at which p is true, and which are most plausible amongst the states at which p is true. Unlike Lewis, Stalnaker assumes that there is always a *unique* most plausible p-state, but apart from this the two approaches are basically equivalent. An important difference between the Grove model and the Lewis–Stalnaker model is that Grove assumes just one global plausibility ordering, whereas Lewis and Stalnaker consider a local plausibility ordering for every state x.

In a *dynamic game*, a player holds at each of his information sets some conditional belief about the opponents' strategies and belief hierarchies. In the previous section we have seen that the players' belief hierarchies can be encoded by means of types within a epistemic model $M = (T_i, b_i)_{i \in I}$. Moreover, M is guaranteed to capture, for every player i, *all* possible conditional belief vectors on $S_{-i} \times T_{-i}$ if we require M to be *belief complete*. So, if we take a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$, then the space of uncertainty for player i is given by $S_{-i} \times T_{-i}$ – the set of all opponents' strategy-type combinations. Consequently, a plausibility ordering for player i is an ordering over the set $S_{-i} \times T_{-i}$.

Definition 5.1 (Plausibility ordering) *Consider a dynamic game G and a belief complete epistemic model $M = (T_i, b_i)$ for G. Then, a plausibility ordering for player i is a binary relation $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$ that is*

(a) total, i.e. for every two strategy-type combinations x and y in $S_{-i} \times T_{-i}$ either $x \succcurlyeq_i y$ or $y \succcurlyeq_i x$,
(b) reflexive, i.e. $x \succcurlyeq_i x$ for every $x \in S_{-i} \times T_{-i}$, and
(c) transitive, i.e. $x \succcurlyeq_i z$ whenever there is some y with $x \succcurlyeq_i y$ and $y \succcurlyeq_i z$.

The meaning of $x \succcurlyeq_i y$ is that player i deems the opponents' strategy-type combination x *at least as plausible* as y. Consider now a player i who holds a plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$, and who observes that his information set h has been reached. Then, player i *knows* that the opponents' strategy-type combination must be somewhere in $S_{-i}(h) \times T_{-i}$, as $S_{-i}(h)$ contains precisely those opponents' strategy combinations that make reaching h possible. If player i's belief revision policy is governed by his plausibility ordering $\succcurlyeq_i$, then player i should concentrate his conditional belief at h on those strategy-type combinations in $S_{-i}(h) \times T_{-i}$ that he deems *most plausible.* That is, player i must concentrate his conditional belief on the set

$$\max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i}) := \{x \in S_{-i}(h) \times T_{-i} \mid x \succcurlyeq_i y \text{ for all } y \in S_{-i}(h) \times T_{-i}\}.$$

But for this conditional belief to be well-defined, we must require that the set $\max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i})$ is non-empty. So, we must require that at each of player i's information sets, there is at least one most plausible strategy-type combination in $S_{-i}(h) \times T_{-i}$. This condition is not automatically satisfied as the set T_{-i} is infinite – in fact uncountably infinite – whenever the epistemic model is belief complete and the game is non-trivial. A plausibility ordering that satisfies this additional requirement is called *well-ordered.*

Definition 5.2 (Well-ordered) *A plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$ is well-ordered if the set $\max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i})$ is non-empty for every information set $h \in H_i$.*

It turns out that there is a close connection between well-ordered plausibility orderings and *systems of spheres* as used in Grove (1988). Namely, for a given well-ordered plausibility ordering $\succcurlyeq_i$, consider for every $x \in S_{-i} \times T_{-i}$ the set

$$sphere_x := \{y \in S_{-i} \times T_{-i} \mid y \succcurlyeq_i x\}.$$

Then, the collection of sets $\{sphere_x \mid x \in S_{-i} \times T_{-i}\}$ is nested, that is, either $sphere_x \subseteq sphere_y$ or $sphere_y \subseteq sphere_x$ for all x, y. Moreover, the well-ordering condition guarantees that for every information set $h \in H_i$ there is a smallest sphere in the collection that intersects $S_{-i}(h) \times T_{-i}$.

Hence, the collection $\{sphere_x \mid x \in S_{-i} \times T_{-i}\}$ corresponds to a *system of spheres* as in Grove (1988). The other direction is also true: If we start from a Grovean system of spheres on $S_{-i} \times T_{-i}$, then this naturally induces a well-ordered plausibility ordering on $S_{-i} \times T_{-i}$. We may therefore interchangeably speak about well-ordered plausibility orderings and systems of spheres – both ways of modelling are equivalent.

5.2 Unique Plausibility Orderings for Reasoning Contexts

Consider a dynamic G and a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$. Then, every type $t_i \in T_i$ holds at every information set $h \in H_i^*$ a conditional belief $b_i(t_i, h) \in \Delta(S_{-i}(h) \times T_{-i})$ on the space of uncertainty $S_{-i} \times T_{-i}$. In particular, the support of $b_i(t_i, h)$ – which we denote by supp $b_i(t_i, h)$ – represents the set of opponents' strategy-type pairs that t_i deems possible at h.

Now, fix a well-ordered plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$. Then, we say that the type *respects* the plausibility ordering $\succcurlyeq_i$ if at every information set $h \in H_i^*$, the type t_i only deems possible strategy-type pairs that are deemed most plausible at h by $\succcurlyeq_i$. We can formally state this as follows.

Definition 5.3 (Type respecting a plausibity ordering) *Consider a dynamic game G and a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$. For a given player i, consider a type $t_i \in T_i$ and a well-ordered plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$. Then, type t_i respects the plausibility ordering $\succcurlyeq_i$ if*

$$\text{supp } b_i(t_i, h) \subseteq \max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i})$$

at every information set $h \in H_i^*$.

In a sense, the plausibility ordering $\succcurlyeq_i$ imposes at every information set $h \in H_i^*$ an upper bound – $\max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i})$ – on the (support of the) conditional beliefs that can be held there. In terms of sphere systems, the above definition states that at every information set h the type t_i looks for the smallest sphere A that intersects $S_{-i}(h) \times T_{-i}$, and concentrates at h on the intersection of $S_{-i}(h) \times T_{-i}$ with A. This is diagrammatically represented in Figure 2, where A is the sphere with the thick border.

Consider next a reasoning context ρ, which selects for the dynamic game G and the epistemic model M some subset of types $\rho_i(G, M) \subseteq T_i$ for player i. That is, the reasoning context ρ puts some restrictions on player i's belief hierarchies in G. The question we are interested in is whether these restrictions can be *characterized* by a *unique* plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$. So, can we find a single plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$ such that the reasoning context ρ selects for player i *precisely* those types t_i that respect $\succcurlyeq_i$?

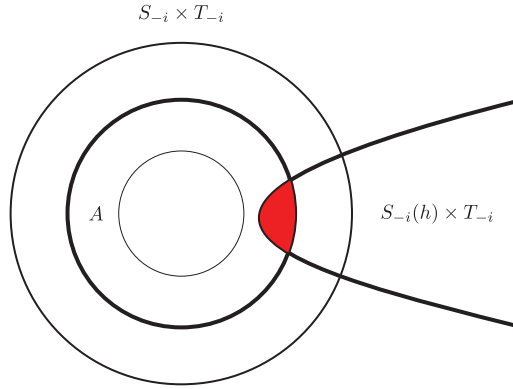

FIGURE 2. (Colour online) Type respecting a plausibility ordering

Definition 5.4 (Reasoning context characterized by plausibility ordering) *Consider a dynamic G, a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$ and a reasoning context ρ, selecting for every player i some subset of types $\rho_i(G, M) \subseteq T_i$. For every player i, consider a well-ordered plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$. Then, the reasoning context is characterized at (G, M) by the profile $(\succcurlyeq_i)_{i \in I}$ of plausibility orderings if for every player i,*

$$\rho_i(G, M) = \{t_i \in T_i \mid t_i \text{ respects } \succcurlyeq_i\}.$$

Hence, if we know the single plausibility ordering $\succcurlyeq_i$ for player i, then we also know precisely which belief hierarchies are selected for player i by the reasoning context.

5.3 Discussion

The definition above can be decomposed into two separate parts. The first part states that all player i belief hierarchies selected by the reasoning context ρ should respect the *same* plausibility ordering $\succcurlyeq_i$. That is, at every information set $h \in H_i^*$ there is a *common* upper bound – $\max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i})$ – for the (supports of) *all* conditional beliefs selected by ρ. This condition alone is not very restrictive, however. What one can always do is to take the trivial plausibility ordering $\succcurlyeq_i^{trivial}$, which deems all opponents' strategy-type combinations as equally plausible, and which has the property that $\max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i}) = S_{-i}(h) \times T_{-i}$ for every information set $h \in H_i^*$. Then, it is trivially true that every type $t_i \in T_i$ respects $\succcurlyeq_i^{trivial}$.

The second part requires, in turn, that *every* type $t_i \in T_i$ which respects the plausibility ordering $\succcurlyeq_i$ *must necessarily be selected* by the reasoning context ρ. So, not only does $\succcurlyeq_i$ impose, at every information set $h \in H_i^*$,

a common upper bound – $\max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i})$ – on the (supports of) all conditional beliefs selected by ρ, but these upper bounds are also *sharp.* By the latter we mean that the reasoning context ρ will select at least one type t_i for which

$$\text{supp } b_i(t_i, h) = \max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i})$$

at all information sets $h \in H_i^*$. That is, the bounds imposed by the plausibility ordering $\succcurlyeq_i$ will actually be covered by some of the belief hierarchies selected by the reasoning context ρ.

By combining these two parts we make sure that the *full* set of belief hierarchies for player i selected by ρ can actually be *characterized* by *one and the same* plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$. Not only do all belief hierarchies selected by ρ respect the same, common plausibility ordering $\succcurlyeq_i$, but also all belief hierarchies that *do* respect this plausibility ordering $\succcurlyeq_i$ are actually selected by ρ. It is thus justified to say that the reasoning context ρ is *characterized* by $\succcurlyeq_i$.

A similar condition could be stated for *individual* types or belief hierarchies. For a *single* type $t_i \in T_i$ and plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$, we could say that t_i is *qualitatively characterized* by $\succcurlyeq_i$ if

$$\text{supp } b_i(t_i, h) = \max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i})$$

for all information sets $h \in H_i^*$. But by Grove's (1988) theorem this would be equivalent to stating that for type t_i, the induced qualitative (non-probabilistic) conditional beliefs, supp $b_i(t_i, h)$, must satisfy the AGM-axioms. In particular, every type t_i that satisfies Bayesian updating whenever possible, can always be qualitatively characterized by some plausibility ordering $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$.

But what we require in Definition 5.4 goes much beyond the AGM-axioms, or Bayesian updating. Instead of requiring that every *individual* type can be characterized by an *individual* plausibility ordering, we impose that the *full* set of selected types can be characterized by a *common* plausibility ordering that applies to *all types* simultaneously.

6. COMMON STRONG BELIEF IN RATIONALITY

6.1 Definition

The reasoning context of 'common strong belief in rationality' has been developed by Battigalli and Siniscalchi (2002). They have shown that the strategies that can rationally be chosen by players who reason in accordance with this concept correspond precisely to the *extensive form rationalizable* strategies as defined by Pearce (1984) and Battigalli (1997). The main idea behind 'common strong belief in rationality' is that a player must believe in the opponents' rationality whenever this is possible. More

precisely, if player i finds himself at information set h, and concludes that h *could* be reached if his opponents choose rationally, then player i *must* believe at h that his opponents choose rationally. We say that player i *strongly believes* in the opponents' rationality. Moreover, if h could be reached if his opponents choose rationally, then player i asks a second question, namely whether h could still be reached if his opponents do not only choose rationally but also strongly believe in their opponents' rationality. If the answer is yes, then player i *must* believe at h that his opponents choose rationally *and* strongly believe in their opponents' rationality. By iterating this argument, we arrive at 'common strong belief in rationality'. To formalize this notion, let us first define what we mean by *rationality* and *strong belief.*

Consider a type t_i for player i, an information set $h \in H_i$ and a strategy s_i that possibly reaches 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 opponents' strategy-type combinations. Note that this expected utility does not depend on the full conditional belief that t_i holds at h, but only on the conditional belief about the opponents' strategy choices.

Definition 6.1 (Rational choice) *Consider a type t_i for player i, an information set $h \in H_i$ and a strategy s_i that possibly reaches h. Strategy s_i is rational for type t_i at information set h if $u_i(s_i, b_i(t_i, h)) \geq u_i(s_i', b_i(t_i, h))$ for all alternative strategies s_i' that possibly reach h. Strategy s_i is rational for type t_i if it is so at every information set $h \in H_i$ that s_i possibly reaches.*

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

Definition 6.2 (Strong belief) *Consider a type t_i within a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$, and an event $E \subseteq S_{-i} \times T_{-i}$. Type t_i strongly believes the event E if $b_i(t_i, h)(E) = 1$ at every information set $h \in H_i^*$ where $(S_{-i}(h) \cap T_{-i}) \cap E$ is non-empty.*

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

Definition 6.3 (Common strong belief in rationality) *Consider a dynamic game G and a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$. For every player i we recursively define sets T_i^k and R_i^k as follows.*

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

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

$$T_i^k := \{t_i \in T_i^{k-1} \mid t_i \text{ strongly believes } R_{-i}^{k-1}\}, \text{ and}$$

$$R_i^k := \{(s_i, t_i) \in S_i \times T_i^k \mid s_i \text{ rational for } t_i\}.$$

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

Here, R_{-i}^{k-1} denotes the set $\times_{j \in I \setminus \{i\}} R_j^{k-1}$. We say that a type t_i expresses 'common strong belief in rationality' if $t_i \in T_i^\infty$. Battigalli and Siniscalchi (2002) show that the sets of types T_i^∞ are always non-empty for every finite dynamic game, and that the strategies which are optimal for a type in T_i^∞ are precisely the *extensive form rationalizable* strategies as defined in Pearce (1984) and Battigalli (1997). By construction, the sets of types T_i^k are monotonically shrinking in k, that is, $T_i^{k+1} \subseteq T_i^k$ for every k. It turns out, actually, that typically these sets will be *strictly* shrinking *for every* k. That is, typically $T_i^{k+1} \subset T_i^k$ for every k, where $\subset$ means strict set inclusion.[3] However, the intersection of all these sets – which is T_i^∞ – will always be non-empty.

6.2 Characterization Result

We will now prove that the reasoning context of 'common strong belief in rationality' *can* be characterized by a unique plausibility ordering for every player, and show how such plausibility orderings can be defined.

Theorem 6.4 (Characterization by plausibility orderings) *Consider a dynamic game G and belief complete epistemic model $M = (T_i, b_i)_{i \in I}$. For every player i consider the binary relation $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$ given by*

$$(s_{-i}, t_{-i}) \succcurlyeq_i (s'_{-i}, t'_{-i}) \text{ if for every } k \in \{0, 1, ...\}:$$

$$(s_{-i}, t_{-i}) \in R_{-i}^k \text{ whenever } (s'_{-i}, t'_{-i}) \in R_{-i}^k .$$

Then, $\succcurlyeq_i$ is a well-ordered plausibility ordering for every player i, and 'common strong belief in rationality' is characterized at (G, M) by the profile $(\succcurlyeq_i)_{i \in I}$ of plausibility orderings.

The proof can be found in the appendix. As the proof of the theorem above shows, the concept of 'common strong belief in rationality' can alternatively be characterized by the Grovean system of spheres

$$R_{-i}^{-1} \supseteq R_{-i}^0 \supseteq R_{-i}^1 \supseteq ... \supseteq R_{-i}^\infty$$

for every player i, where we set $R_{-i}^{-1} := S_{-i} \times T_{-i}$. That is, player i will look at every information set $h \in H_i$ for the smallest sphere R_{-i}^k that

[3] I am grateful to Amanda Friendenberg who pointed this out to me.

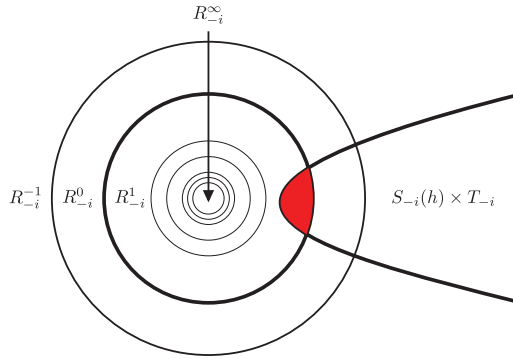

FIGURE 3. (Colour online) Characterization of 'common strong belief in rationality' by Grovean system of spheres

intersects $S_{-i}(h) \times T_{-i}$, and will concentrate his conditional belief at h on the intersection between $S_{-i}(h) \times T_{-i}$ and this smallest sphere R_{-i}^{k}. This is diagrammatically represented in Figure 3.

In this picture, we have taken R_{-i}^{1} to be the smallest sphere that intersects $S_{-i}(h) \times T_{-i}$.

7. COMMON BELIEF IN FUTURE RATIONALITY

7.1 Definition

The concept of 'common belief in future rationality' has been defined in Perea (2014), and is very similar to the notion of 'common knowledge of stable belief in dynamic rationality' by Baltag *et al.* (2009). The main difference between the two is that the latter notion restricts to dynamic games with perfect information whereas the first is applicable to all finite dynamic games. The key idea is that a player must always believe, at every stage of the game, that his opponents will choose rationally in the game that lies ahead. We say that this player *believes in his opponents' future rationality.* Baltag *et al.* (2009) refer to this condition as 'stable belief in dynamic rationality'. Not only this, a player must also always believe that his opponents always believe in their opponents' future rationality, and so on. This eventually leads to the concept of 'common belief in future rationality'.

This concept is completely *forward looking*, as a player need not necessarily believe that his opponents have chosen rationally in the *past* even when believing so is possible. At the same time, a player must always hold on to the belief that his opponents will choose rationally in the future even when it is evident that these same opponents have chosen irrationally in the past. So, in a sense, it requires a degree of 'stubbornness'

by the players that is not present in 'common strong belief in rationality'. To formally define the concept, we first state precisely what we mean by 'belief in future rationality'.

For a given strategy s_i for player i, let $H_i(s_i)$ denote the collection of information sets for player i that are possibly reached by s_i. Consider a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$ for the dynamic game G at hand. For every information set h in the game, let

$$R_i[h] := \{(s_i, t_i) \in S_i \times T_i \mid s_i \text{ is rational for } t_i \text{ at every } h' \in H_i(s_i) \text{ weakly following } h\}.$$

Remember that h' weakly follows h if either h' follows h, or h' and h are simultaneous. Hence, $R_i[h]$ contains those strategy-type pairs where the strategy is optimal for the type 'from h onwards'.

Definition 7.1 (Belief in future rationality) *A type t_i believes in his opponents' future rationality if at every information set $h \in H_i^*$, the conditional belief $b_i(t_i, h)$ assigns probability 1 to the event $R_{-i}[h]$.*

Here, $R_{-i}[h] := \times_{j \in I \setminus \{i\}} R_j[h]$. So, no matter what has happened in the game so far, type t_i will always at every information set h assign probability 1 to the event that his opponents will choose rationally from h onwards. With this definition at hand, we can now formally introduce 'common belief in future rationality'.

Definition 7.2 (Common belief in future rationality) *Consider a dynamic game G and a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$. For every player i we recursively define sets T_i^k as follows.*

Induction start. *Define $T_i^1 := \{t_i \in T_i \mid t_i$ believes in his opponents' future rationality $\}$.*

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

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

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

Hence, a type in T_i^k always believes that every opponent j holds a type in T_j^{k-1}. In Perea (2014) it is shown that the set T_i^∞ is always non-empty for every finite dynamic game. Moreover, both Perea (2014) and Baltag *et al.* (2009) show that in every dynamic game with perfect information without relevant ties, the strategies selected by the concept are precisely the backward induction strategies.

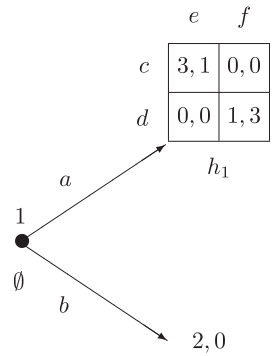

FIGURE 4. A game in which 'common belief in future rationality' cannot be characterized by plausibility orderings

7.2 Impossibility Result

We show that the reasoning context of 'common belief in future rationality' can *not* always be characterized by a unique plausibility ordering for every player. Consider the game G in Figure 4, which is known as 'Battle-of-the-sexes-with-outside-option' and constitutes one of the classical forward induction examples in the literature.

Take an arbitrary belief complete epistemic model $M = (T_i, b_i)_{i \in I}$. We show that the reasoning context of 'common belief in future rationality' cannot be characterized at (G, M) by any profile of plausibility orderings.

We first show that there must be a type $t_2^* \in T_2$ for player 2 that expresses 'common belief in future rationality', and which initially believes that player 1 chooses (a, c). Namely, as the model M is belief complete, there must be types $t_1^* \in T_1$ and $t_2^* \in T_2$ with the following conditional beliefs:

$$b_1(t_1^*, \emptyset) = (e, t_2^*), \qquad b_1(t_1^*, h_1) = (e, t_2^*)$$
$$b_2(t_2^*, \emptyset) = ((a, c), t_1^*), \; b_2(t_2^*, h_1) = ((a, c), t_1^*).$$

Here, $b_1(t_1^*, \emptyset) = (e, t_2^*)$ means that type t_1^* ascribes at $\emptyset$ probability 1 to the event that player 2 chooses strategy e while being of type t_2^*. Similarly for the other three beliefs.

It can easily be verified that both types t_1^* and t_2^* believe in the opponent's future rationality. Consider, for instance, the type t_2^*. That type believes at $\emptyset$ that player 1 chooses (a, c) and that player 1 is of type t_1^*. As type t_1^* believes, at $\emptyset$ and h_1, that player 2 chooses e, strategy (a, c) is optimal for t_1^* at $\emptyset$ and h_1. Hence, type t_2^* believes at $\emptyset$ that player 1 chooses optimally at $\emptyset$ and h_1, so t_2^* believes at $\emptyset$ in 1's future rationality. Similarly, it can be checked that the same type t_2^* believes at h_1 that player 1 chooses optimally at h_1. Therefore, t_2^* believes in 1's future rationality overall. In

the same way it can be checked that type t_1^* also believes in his opponent's future rationality. As t_1^* believes throughout the game that player 2's type is t_2^*, and t_2^* believes throughout the game that player 1's type is t_1^*, it immediately follows that both t_1^* and t_2^* express 'common belief in future rationality'. In particular, $t_2^* \in T_2$ is a type that expresses 'common belief in future rationality', and which initially believes that player 1 chooses (a, c).

On the other hand, (a, d) can never be an optimal strategy for player 1 at the beginning, as choosing b always yields him a strictly better outcome. So, under 'common belief in future rationality', player 2 cannot initially ascribe positive probability to player 1 choosing (a, d). Consequently, there is no type $t_2 \in T_2$ that expresses 'common belief in future rationality' and that initially assigns positive probability to player 1 choosing (a, d).

Now suppose, contrary to what we want to show, that the concept of 'common belief in future rationality' is characterized at (G, M) by a profile $(\succcurlyeq_i)_{i \in I}$ of well-ordered plausibility orderings. Then, in particular,

$$T_2^\infty = \{t_2 \in T_2 \mid t_2 \text{ respects } \succcurlyeq_2\}, \tag{1}$$

where T_2^∞ is the set of player 2 types in T_2 that express 'common belief in future rationality'. This implies that

$$\text{supp } b_2(t_2, \emptyset) \subseteq \max_{\succcurlyeq_2}(S_1(\emptyset) \times T_1)$$

for all $t_2 \in T_2^\infty$. Obviously, $S_1(\emptyset) = S_1$, so we have that

$$\text{supp } b_2(t_2, \emptyset) \subseteq \max_{\succcurlyeq_2}(S_1 \times T_1)$$

for all $t_2 \in T_2^\infty$. As the type t_2^* above is in T_2^∞, and $b_2(t_2^*, \emptyset) = ((a, c), t_1^*)$, it follows that

$$((a, c), t_1^*) \in \max_{\succcurlyeq_2}(S_1 \times T_1). \tag{2}$$

On the other hand, we have seen that there is no $t_2 \in T_2^\infty$ that initially assigns positive probability to player 1 choosing (a, d). Hence, by (1)

$$((a, d), t_1) \notin \max_{\succcurlyeq_2}(S_1 \times T_1) \text{ for all } t_1 \in T_1. \tag{3}$$

Note that $S_1(h_1) = \{(a, c), (a, d)\}$. By (2) and (3) it then follows that

$$((a, d), t_1) \notin \max_{\succcurlyeq_2}(S_1(h_1) \times T_1) \text{ for all } t_1 \in T_1. \tag{4}$$

We will now show, however, that there is some $\hat{t}_2 \in T_2^\infty$ which at h_1 assigns probability 1 to player 1 choosing (a, d). As the model M is belief complete, there must be types $\hat{t}_1 \in T_1$ and $\hat{t}_2 \in T_2$ with the following conditional beliefs:

$$b_1(\hat{t}_1, \emptyset) = (f, \hat{t}_2),\ b_1(\hat{t}_1, h_1) = (f, \hat{t}_2)$$
$$b_2(\hat{t}_2, \emptyset) = (b, \hat{t}_1),\ b_2(\hat{t}_2, h_1) = ((a, d), \hat{t}_1).$$

Note that type $\hat{t}_2$ revises his belief about player 1's strategy choice during the game: at the beginning, $\hat{t}_2$ believes that player 1 chooses b, whereas at h_1 type $\hat{t}_2$ believes that player 1 chooses (a,d).

It may be verified that both types $\hat{t}_1$ and $\hat{t}_2$ believe in the opponent's future rationality. Consider, for instance, the type $\hat{t}_2$, which believes at $\emptyset$ that player 1 chooses b while being of type $\hat{t}_1$. As type $\hat{t}_1$ believes at $\emptyset$ that player 2 chooses f, strategy b is optimal for type $\hat{t}_1$ at $\emptyset$. Hence, $\hat{t}_2$ believes at $\emptyset$ that player 1 chooses rationally at $\emptyset$. Strategy b for player 1 makes reaching h_1 impossible, so we conclude that type $\hat{t}_2$ believes at $\emptyset$ in 1's future rationality. At h_1, the same type $\hat{t}_2$ believes that player 1 chooses (a,d) while being of type $\hat{t}_1$. As type $\hat{t}_1$ believes at h_1 that player 2 chooses f, strategy (a,d) is optimal for type $\hat{t}_1$ at h_1. Indeed, among the two strategies for player 1 that reach h_1 – which are (a,c) and (a,d) – strategy (a,d) is optimal under the belief that player 2 chooses f. So, type $\hat{t}_2$ believes at h_1 that player 1 chooses rationally at h_1. Overall, we may conclude that type $\hat{t}_2$ believes at $\emptyset$ and h_1 in 1's future rationality. That is, $\hat{t}_2$ believes in 1's future rationality. Note, however, that $\hat{t}_2$ believes at h_1 that player 1 has chosen *irrationally in the past*, as b is better than (a,d) for $\hat{t}_1$ *at the beginning*. This is not a problem, as 'common belief in future rationality' only requires players to believe in the opponents' *future* rationality, not necessarily in the opponents' *past* rationality.

In a similar fashion, it may be verified that also type $\hat{t}_1$ believes in his opponent's future rationality. As $\hat{t}_1$ believes throughout that player 2 is of type $\hat{t}_2$, and $\hat{t}_2$ believes throughout that player 1 is of type $\hat{t}_1$, it follows that both $\hat{t}_1$ and $\hat{t}_2$ express 'common belief in future rationality'. In particular, we have found a type $\hat{t}_2 \in T_2^\infty$ which at h_1 assigns probability 1 to player 1 choosing (a,d).

But then, by (4), it follows that

$$\text{supp } b_2(\hat{t}_2, h_1) \not\subseteq \max_{\succcurlyeq_2}(S_1(h_1) \times T_1).$$

Hence, $\hat{t}_2$ does not respect the plausibility ordering $\succcurlyeq_2$, which contradicts the assumption (1). We are therefore led to conclude that the concept of 'common belief in future rationality' cannot be characterized at (G,M) by a unique plausibility ordering for every player.

So, in a nutshell, the reason why 'common belief in future rationality' cannot be characterized by plausibility orderings in the game of Figure 4 is as follows. Under 'common belief in future rationality', player 1 can rationally choose (a,c) but not (a,d). Therefore, player 2 types which express 'common belief in future rationality' may initially deem (a,c) possible, but certainly not (a,d). Hence, if 'common belief in future rationality' were to be characterized by a unique plausibility ordering on player 1's strategy-type pairs, then this plausibility ordering must necessarily deem (a,c) more plausible than (a,d). But then, upon reaching h_1, player 2 must necessarily conclude that player 1 did not choose (a,d),

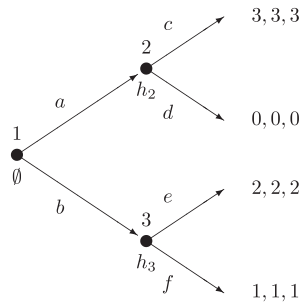

FIGURE 5. A perfect information game in which 'common belief in future rationality' cannot be characterized by plausibility orderings

which is not true since under 'common belief in future rationality' player 2 *can* believe at h_1 that player 1 chooses (a, d).

8. GAMES WITH PERFECT INFORMATION

A dynamic game is said to be with *perfect information* if different players never choose simultaneously, and every player, when making a choice, always knows exactly what the other players have done so far. Formally this means that at every non-terminal history exactly one player is active, and every information set consists of precisely one non-terminal history. We say that the game is *without relevant ties* (see Battigalli 1997) if for every player i, every information set $h \in H_i$, and every two different terminal histories z, z' following h, it holds that $u_i(z) \neq u_i(z')$. Hence, two different choices for player i always lead to different utilities for that player.

It is well-known that in every perfect information game without relevant ties, the backward induction procedure yields a unique choice $c^{bi}(h)$ at every information set h. We will refer to these choices as the *backward induction choices* in the game. The *backward induction strategy* for player i is the unique strategy s_i^{bi} that selects the backward induction choice $c^{bi}(h)$ at every $h \in H_i$ possibly reached by s_i^{bi}.

In Perea (2014) it is shown that in every perfect information game without relevant ties, the concept of 'common belief in future rationality' uniquely selects the backward induction strategy for every player. Indeed, in such games there is only one strategy that a player can rationally choose if his belief hierarchy expresses 'common belief in future rationality', namely his backward induction strategy.

However, even for such games the concept of 'common belief in future rationality' may not be characterizable by plausibility orderings, as the game in Figure 5 shows.

Clearly, under 'common belief in future rationality', player 2 must believe at $\emptyset$ that player 1 believes at $\emptyset$ that (a) player 2 will choose c and (b) player 3 will choose e. Also, player 2 must believe at $\emptyset$ that player 1 chooses rationally at $\emptyset$ and that player 3 chooses rationally at h_3. As such, under 'common belief in future rationality' player 2 must believe at $\emptyset$ that player 1 will choose a and that player 3 would choose e at h_3. If 'common belief in future rationality' were to be characterized by a unique plausibility ordering $\succcurlyeq_2$ for player 2, then $\succcurlyeq_2$ must deem the strategy combination (a, e) most plausible overall. But then, player 2 should still believe at h_2 that player 1 has chosen a and that player 3 would have chosen e at h_3. However, under 'common belief in future rationality', player 2 is free to believe at h_2 that player 3 would have chosen f, as player 3's information set h_3 does not follow h_2. Hence, in this game with perfect information, 'common belief in future rationality' cannot be characterized by a unique plausibility ordering for player 2.

At the same time, the example in Figure 5 shows that 'common belief in future rationality' is perhaps a bit too permissive. Indeed, there is no good reason why player 2 at h_2 should suddenly drop his belief that player 3 would choose rationally at h_3. This leads to the question whether we can strengthen the concept of 'common belief in future rationality' such that the new, more restrictive concept *can* be characterized by plausibility orderings in perfect information games without relevant ties. We will see that this is indeed possible.

Instead of only requiring a player to believe that his opponents will choose rationally at *future* information sets, let us look at a stronger condition which states that at any point in time, a player also believes that his opponents *would* have chosen rationally at information sets that have been *avoided by past choices.* We call such information sets *parallel* information sets. For instance, in Figure 5 the information set h_3 is parallel to information set h_2 as it is avoided by the past choice a that leads to h_2. A more formal way of stating it is to say that an information set h' is *parallel* to another information set h if h' does not weakly precede, nor weakly follow, h.

The condition above, that a player always believes that his opponents will choose rationally in the future, and would have chosen rationally at parallel information sets, can formally be stated as follows. Consider a dynamic G – not necessarily with perfect information – and a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$ for G. For every player i, and every information set h in G, define the event

$$\hat{R}_i[h] := \{(s_i, t_i) \in S_i \times T_i \mid s_i \text{ is rational for } t_i \text{ at every } h' \in H_i(s_i) \text{ weakly following } h, \text{and every } h' \in H_i(s_i) \text{ that is parallel to } h\}.$$

Definition 8.1 ***(Belief in rationality at future and parallel information sets)*** *A type t_i believes in his opponents' rationality at future and parallel information sets if at every information set $h \in H_i^*$, the conditional belief $b_i(t_i, h)$ assigns probability 1 to the event $\hat{R}_{-i}[h]$.*

Here, $\hat{R}_{-i}[h] := \times_{j \in I \setminus \{i\}} \hat{R}_j[h]$. With this basic condition at hand, the reasoning context of 'common belief in rationality at future and parallel information sets' can then be defined in the obvious way.

Definition 8.2 ***(Common belief in rationality at future and parallel information sets)*** *Consider a dynamic game G and a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$. For every player i we recursively define sets T_i^k as follows.*

Induction start. *Define $T_i^1 := \{t_i \in T_i \mid t_i$ believes in his opponents' rationality at future and parallel information sets $\}$.*

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

$$T_i^k := \left\{ t_i \in T_i^{k-1} \mid b_i(t_i, h)\left(S_{-i} \times T_{-i}^{k-1}\right) = 1 \text{ for all } h \in H_i^* \right\}.$$

Common belief in rationality at future and parallel information sets selects for every player i the set of types $T_i^\infty := \cap_{k \in \mathbb{N}} T_i^k$.

It can be shown that the sets of types T_i^∞ that express 'common belief in rationality at future and parallel information sets' will always be non-empty, and will always be included in the sets of types that express 'common belief in future rationality'. However, the two concepts are 'behaviourally equivalent' as they always select the same sets of strategies for every player. The reason is that for player i's choice at information set h it is only relevant what player i believes about the opponents' *past* and *future* choices, not what he believes about the opponents' possible behaviour at parallel information sets. Clearly, the two concepts above impose no conditions on player i's belief about his opponents' past choices, and impose exactly the same conditions on his belief about the opponents' future choices. As such, it does not matter for player i's choice at information set h whether his beliefs are restricted by 'common belief in rationality at future and parallel information sets' or only by 'common belief in future rationality'. Therefore, the two concepts only differ in the restrictions they impose on the players' conditional beliefs, but not in the strategy choices they select for the players. The formal proofs for the insights above are not difficult, and we leave these to the reader for the sake of brevity.

As discussed above, the concept of 'common belief in future rationality' uniquely filters the backward induction strategies for every perfect information game without relevant ties. It then immediately follows also that 'common belief in rationality at future and parallel information sets' uniquely selects the backward induction strategies in

such games, as it is behaviourally equivalent to 'common belief in future rationality'. We can actually say a little more: Under 'common belief in rationality at future and parallel information sets', there will be a *unique* belief for every player i at each of his information sets $h \in H_i^*$ about the opponents' strategy choices, namely that his opponents will choose the *backward induction choices* at all *future* and *parallel* information sets. This is not true for 'common belief in future rationality'. In the game of Figure 5, for instance, player 2 may believe at h_2 under 'common belief in future rationality' that player 3 would choose f at h_3, although f is not the backward induction choice at h_3.

Formally speaking, for a given player i and information set h in the game, let $s_i^{bi}[h]$ be the unique strategy that (a) at every $h' \in H_i$ preceding h selects the unique choice leading to h, and (b) at every $h' \in H_i$ not preceding h selects the backward induction choice $c^{bi}(h')$ whenever h' is possibly reached by $s_i^{bi}[h]$. So, $s_i^{bi}[h]$ possibly reaches h, and selects the backward induction choices at all future and parallel information sets to h. We call $s_i^{bi}[h]$ the backward induction strategy *conditional on* h. For a player i and information set $h \in H_i^*$, let $s_{-i}^{bi}[h] := (s_j^{bi}[h])_{j \in I \setminus \{i\}}$ be the combination of opponents' backward induction strategies conditional on h.

It can be shown that under 'common belief in rationality at future and parallel information sets', a type for player i must at every $h \in H_i^*$ assign probability 1 to the strategy profile $s_{-i}^{bi}[h]$ by the opponents. This proof is not difficult, and is left to the reader. Hence, every player i holds a unique vector β_i^{bi} of conditional beliefs about the opponents' strategy choices. But then, every player i must believe throughout the game that every opponent j holds the belief vector β_j^{bi}, and must believe throughout that every opponent j believes throughout that every other player k holds the conditional belief vector β_k^{bi}, and so on. Clearly, this leads to the conclusion that under 'common belief in rationality at future and parallel information sets', the full belief hierarchy of every player is *uniquely determined.* That is, the set T_i^∞ of types expressing 'common belief in rationality at future and parallel information sets' is a singleton. Let us denote by t_i^{bi} the unique type for player i that expresses 'common belief in rationality at future and parallel information sets'. Then, at every information set $h \in H_i^*$ the conditional belief of type t_i^{bi} about the opponents' strategy-type combinations is given by

$$b_i\left(t_i^{bi}, h\right) = \left(s_{-i}^{bi}[h], t_{-i}^{bi}\right).$$

That is, at $h \in H_i^*$ type t_i^{bi} assigns probability 1 to the event that every opponent j chooses the backward induction strategy $s_j^{bi}[h]$ conditional on h while being of type t_j^{bj}. This insight will be important for proving the following result, which states that for perfect information games without relevant ties, the reasoning context of 'common belief in rationality at

future and parallel information sets' can be characterized by a unique plausibility ordering for every player.

Theorem 8.3 (Characterization by plausibility orderings) *Consider a perfect information game G without relevant ties and a belief complete epistemic model $M = (T_i, b_i)_{i \in I}$. For every player i consider the binary relation $\succcurlyeq_i$ on $S_{-i} \times T_{-i}$ given by: $(s_{-i}, t_{-i}) \succ_i (s'_{-i}, t'_{-i})$ if either*

(a) $t_{-i} = t_{-i}^{bi}$ *and* $t'_{-i} \neq t_{-i}^{bi}$, *or*
(b) $s_{-i} = s_{-i}^{bi}[h]$ *for some* $h \in H_i^*$ *and* $s'_{-i} \neq s_{-i}^{bi}[h]$ *for any* $h \in H_i^*$, *or*
(c) $s_{-i} \neq s'_{-i}$, *and there are some* $h, h' \in H_i^*$, *where* h' *follows* h, *such that* $s_{-i} = s_{-i}^{bi}[h]$ *and* $s'_{-i} = s_{-i}^{bi}[h']$.

Then, $\succcurlyeq_i$ is a well-ordered plausibility ordering for every player i, and 'common belief in rationality at future and parallel information sets' is characterized at (G, M) by the profile $(\succcurlyeq_i)_{i \in I}$ of plausibility orderings.

The proof can be found in the appendix. The theorem above cannot be extended to general dynamic games, however. Consider for instance the game in Figure 4. For that game, the concept of 'common belief in rationality at future and parallel information sets' is fully equivalent – also in terms of beliefs – to 'common belief in future rationality' as there are no parallel information sets in that game. Hence, even the concept of 'common belief in rationality at future and parallel information sets' cannot be characterized by plausibility orderings in that game.

The theorem above shows, in particular, that for perfect information games without relevant ties we can always find a reasoning context that (a) uniquely selects the backward induction strategy for every player, and (b) can be characterized by plausibility orderings. In that respect, the result is very similar to Baltag *et al.* (2009). These authors, namely, *assume* from the beginning that the conditional beliefs of the players are characterized by a unique plausibility ordering for every player. Based on this assumption they then derive the notion of 'common knowledge of stable belief in dynamic rationality', which is very similar to 'common belief in future rationality', but now with the additional assumption that conditional beliefs are derived from plausibility orderings. In Corollary 4.5 they then prove that the concept of 'common knowledge of stable belief in dynamic rationality' uniquely yields the backward induction strategies in every perfect information game without relevant ties. So Baltag *et al.* (2009) also present a reasoning context that is characterized by plausibility orderings and that uniquely returns the backward induction strategies in perfect information games without relevant ties.

9. DISCUSSION

We have seen that for the concept of 'common strong belief in rationality', the whole collection of selected belief hierarchies for a given player can be summarized by a *single plausibility ordering*, whereas this is not always possible for 'common belief in future rationality'. Can this be viewed as an argument in favour of the first concept, and against the second? Not necessarily. A player in a dynamic game is in general not interested in finding *all possible* belief hierarchies that he could reasonably hold, relative to a given reasoning context, but rather aims at producing *one* such reasonable belief hierarchy. The property above, that all selected belief hierarchies for a given player can be summarized by a unique plausibility ordering, is therefore of interest mainly to the *game-theorist* or *analyst* – who looks at the game from a meta-perspective – rather than to the players themselves. In fact, there is nothing wrong with the concept of 'common belief in future rationality' – the concept is logically sound and is based on rather intuitive assumptions. Moreover, from a player's perspective the concept is not more complex than 'common strong belief in rationality'. From a meta-perspective, however, 'common strong belief in rationality' can be viewed as somewhat simpler since the entire set of belief hierarchies selected for a given player is characterized by a single plausibility ordering.

But even from a meta-perspective, it is not necessarily true that 'common strong belief in rationality' is more natural than 'common belief in future rationality'. Indeed, why should we necessarily require that all selected belief hierarchies for a given player share the same plausibility ordering? Within the bounds of 'common belief in future rationality', it is often the case that different belief hierarchies for the same player are based on different plausibility orderings. At the same time, these belief hierarchies still share an important common feature, namely that they believe in the opponents' future rationality, believe that the other players believe in their opponents' future rationality, and so on. The difference with 'common strong belief in rationality' is that this common feature cannot be reduced to a common plausibility ordering. But why should this be the case?

The investigation we have carried out in this paper is therefore primarily of a descriptive – and not of a normative – character. We do not make any normative judgements about the concepts of 'common strong belief in rationality' and 'common belief in future rationality' – in fact we believe that both concepts are quite natural, and have their own intuitive appeal.

REFERENCES

Alchourrón, C.E., P. Gärdenfors and D. Makinson. 1985. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic* 50: 510–530.

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

Battigalli, P. 1997. On rationalizability in extensive games. *Journal of Economic Theory* 74: 40–61.

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

Battigalli, P. and M. Siniscalchi. 2002. Strong belief and forward induction reasoning. *Journal of Economic Theory* 106: 356–391.

Bonanno, G. 2009. Revealed preference, iterated belief revision and dynamic games. In *Dagstuhl Seminar Proceedings* 09351, *Information processing, rational belief change and social interaction.*

Bonanno, G. 2011. AGM-belief revision in dynamic games, In *Proceedings of Theoretical Aspects of Rationality and Knowledge (TARK) XIII.*

Bonanno, G. 2013. AGM-consistency and perfect Bayesian equilibrium. Part I: definition and properties. *International Journal of Game Theory* 42: 567–592.

Brandenburger, A. 2003. On the existence of a 'complete' possibility structure. In *Cognitive Processes and Economic Behavior*, ed. N. Dimitri, M. Basili and I. Gilboa. London: Routledge.

Grove, A. 1988. Two modellings for theory change. *Journal of Philosophical Logic* 17: 157–170.

Harsanyi, J. C. 1967–1968. Games with incomplete information played by 'Bayesian' players, I–III. *Management Science* 14: 159–182, 320–334, 486–502.

Levi, I. 2013. How infallible but corrigible full belief is possible. In *A Formal Epistemology Reader*, ed. H. Arló-Costa, V. F. Hendricks and J. van Benthem. Cambridge: Cambridge University Press.

Lewis, D. K. 1973. *Counterfactuals*. Oxford: Blackwell.

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

Perea, A. 2014. Belief in the opponents' future rationality. *Games and Economic Behavior* 84: 231–254.

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

Stalnaker, R. C. 1968. A theory of conditionals. *Studies in Logical Theory*: 98–112.

APPENDIX

Proof of Theorem 6.4. Fix a player i. We first show that the binary relation $\succcurlyeq_i$ defined in the statement of the theorem is total, reflexive, transitive and well-ordered. To prove so, it will be helpful to introduce some additional objects.

Let $R_{-i}^{-1} := S_{-i} \times T_{-i}$, and $R_{-i}^{\infty} := \cap_{k \in \mathbb{N}} R_{-i}^{k}$. Then we have that

$$S_{-i} \times T_{-i} = R_{-i}^{-1} \supseteq R_{-i}^{0} \supseteq R_{-i}^{1} \supseteq ... \supseteq R_{-i}^{\infty}.$$

Define $K := \{-1, 0, 1, ...\} \cup \{\infty\}$. So, the collection $\{R_{-i}^{k} \mid k \in K\}$ of subsets is nested, with R_{-i}^{1} being the full space $S_{-i} \times T_{-i}$. Moreover, R_{-i}^{∞} is non-empty as shown in Battigalli and Siniscalchi (2002).

For every element $x \in S_{-i} \times T_{-i}$, define the number $k(x) := \max\{k \in K \mid x \in R_{-i}^{k}\}$. It is easily seen that $k(x)$ is well-defined. Namely, if $x \in R_{-i}^{\infty}$, then $k(x) = \infty$ by definition. Suppose, on the other hand, that $x \notin R_{-i}^{\infty}$. Since $R_{-i}^{\infty} = \cap_{k \in \mathbb{N}} R_{-i}^{k}$, there must be some $k \in K \backslash \{\infty\}$ such that $x \in R_{-i}^{k}$ but $x \notin R_{-i}^{k+1}$. But then, $k(x) = k$. So, indeed, $k(x)$ is well-defined for every $x \in S_{-i} \times T_{-i}$.

By definition of $\succcurlyeq_i$ we have that

$$x \succcurlyeq_i y \text{ if and only if } k(x) \geq k(y)$$

for all $x, y \in S_{-i} \times T_{-i}$. But then, it is immediately clear that $\succcurlyeq_i$ is total, reflexive and transitive.

It remains to show that $\succcurlyeq_i$ is well-ordered. Consider some information set $h \in H_i^*$. Define the number

$$k(h) := \max\{k \in K \mid R_{-i}^k \cap (S_{-i}(h) \times T_{-i}) \text{ is non-empty}\}.$$

We first show that the number $k(h)$ is well-defined. For every opponent $j \neq i$ and every $k \in K \backslash \{-1\}$, let

$$S_j^k := \{s_j \in S_j \mid s_j \text{ rational for some } t_j \in T_j^k\}.$$

Let $S_{-i}^k := \times_{j \in I \backslash \{i\}} S_j^k$. So, S_{-i}^k is the set of opponents' strategy combinations present in R_{-i}^k. As the dynamic game G is finite, the set S_{-i} of opponents' strategy-combinations is finite as well. But then, there must be some $k^* \in K \backslash \{-1, \infty\}$ such that $S_{-i}^\infty = S_{-i}^{k^*}$. Now, to show that the number $k(h)$ is well-defined we distinguish two cases. Suppose first that $R_{-i}^\infty \cap (S_{-i}(h) \times T_{-i})$ is non-empty. Then, $k(h) = \infty$ by definition. Suppose, on the other hand, that $R_{-i}^\infty \cap (S_{-i}(h) \times T_{-i})$ is empty. Then, $S_{-i}^\infty \cap S_{-i}(h)$ is empty. Since $S_{-i}^\infty = S_{-i}^{k^*}$ we have that $S_{-i}^{k^*} \cap S_{-i}(h)$ is empty, so $R_{-i}^{k^*} \cap (S_{-i}(h) \times T_{-i})$ is empty as well. But then, there must be some $k < k^*$ such that $R_{-i}^k \cap (S_{-i}(h) \times T_{-i})$ is non-empty but $R_{-i}^{k+1} \cap (S_{-i}(h) \times T_{-i})$ is empty. In that case, $k(h) = k$. So, indeed, $k(h)$ is well-defined.

By definition of $\succcurlyeq_i$ we have at information set $h \in H_i^*$ that

$$\max_{\succcurlyeq_i}(S_{-i}(h) \cap T_{-i}) = R_{-i}^{k(h)} \cap (S_{-i}(h) \cap T_{-i}), \tag{5}$$

which is guaranteed to be non-empty. We may therefore conclude that $\succcurlyeq_i$ is well-ordered.

We now show that

$$T_i^\infty = \{t_i \in T_i \mid t_i \text{ respects } \succcurlyeq_i\}, \tag{6}$$

which would imply that 'common strong belief in rationality' is characterized by the unique plausibility ordering $\succcurlyeq_i$ for player i.

For every information set $h \in H_i^*$, remember that $k(h)$ is the highest number $k \in K$ for which $R_{-i}^k \cap (S_{-i}(h) \cap T_{-i})$ is non-empty. By definition of 'common strong belief in rationality', T_i^∞ contains precisely those types $t_i \in T_i$ that strongly believe each of the events R_{-i}^k, $k \in \{0, 1, ...\}$. But then, it follows that

$$\begin{aligned}
T_i^\infty = \{&t_i \in T_i \mid \text{supp } b_i(t_i, h) \subseteq R_{-i}^k \cap (S_{-i}(h) \cap T_{-i}) \\
&\text{whenever } R_{-i}^k \cap (S_{-i}(h) \cap T_{-i}) \neq \emptyset \\
&\text{for all } h \in H_i^*\} \\
= \{&t_i \in T_i \mid \text{supp } b_i(t_i, h) \subseteq R_{-i}^{k(h)} \cap (S_{-i}(h) \cap T_{-i}) \\
&\quad \text{for all } h \in H_i^*\}.
\end{aligned}$$

By (5), we obtain that

$$T_i^\infty = \{t_i \in T_i \mid \text{supp } b_i(t_i, h) \subseteq \max_{\succcurlyeq_i}(S_{-i}(h) \cap T_{-i}) \text{ for all } h \in H_i^*\}.$$

Hence, (6) must hold. So, we conclude that 'common strong belief in rationality' is characterized at (G, M) by $\succcurlyeq_i$ for player i. This holds for every player i, and hence the proof is complete. ■

Proof of Theorem 8.3. Let $M = (T_i, b_i)_{i \in I}$ be a belief complete epistemic model for G. First of all, it can easily be verified that the binary relation $\succcurlyeq_i$ defined above is a well-ordered plausibility ordering. We leave this to the reader. Now, let ρ be the reasoning context of 'common belief in rationality at future and parallel information sets'. As we have seen above, there is for every player i a unique type t_i^{bi} in $\rho_i(G, M)$, and at every information set $h \in H_i^*$ this type holds the conditional belief $b_i(t_i^{bi}, h) = (s_{-i}^{bi}[h], t_{-i}^{bi})$. So, in order to prove that ρ is characterized by the plausibility orderings above, we must show that

$$\max_{\succcurlyeq_i}(S_{-i}(h) \times T_{-i}) = \{(s_{-i}^{bi}[h], t_{-i}^{bi})\} \tag{7}$$

for every player i and every $h \in H_i^*$.

Fix a player i and some information set $h \in H_i^*$. As, by construction, $s_{-i}^{bi}[h] \in S_{-i}(h)$, it follows that $(s_{-i}^{bi}[h], t_{-i}^{bi}) \in S_{-i}(h) \times T_{-i}$. Hence, it remains to show that

$$(s_{-i}^{bi}[h], t_{-i}^{bi}) \succ_i (s_{-i}, t_{-i}) \text{ for all } (s_{-i}, t_{-i}) \in S_{-i}(h) \times T_{-i} \text{ with } (s_{-i}, t_{-i}) \neq (s_{-i}^{bi}[h], t_{-i}^{bi}). \tag{8}$$

Take some arbitrary $(s_{-i}, t_{-i}) \in S_{-i}(h) \times T_{-i}$ with $(s_{-i}, t_{-i}) \neq (s_{-i}^{bi}[h], t_{-i}^{bi})$. We distinguish the following cases.

Case 1. If $t_{-i} \neq t_{-i}^{bi}$ then, by (a) in the statement of the theorem, $(s_{-i}^{bi}[h], t_{-i}^{bi}) \succ_i (s_{-i}, t_{-i})$.

Case 2. If $s_{-i} \neq s_{-i}^{bi}[h']$ for any $h' \in H_i^*$ then, by (b) in the statement of the theorem, $(s_{-i}^{bi}[h], t_{-i}^{bi}) \succ_i (s_{-i}, t_{-i})$.

So, from now on we assume that $t_{-i} = t_{-i}^{bi}$ and $s_{-i} = s_{-i}^{bi}[h']$ for some $h' \in H_i^*$. That is, $(s_{-i}, t_{-i}) = (s_{-i}^{bi}[h'], t_{-i}^{bi})$. We continue by distinguishing the following cases.

Case 3. If h' follows h then, by (c) in the statement of the theorem, $(s_{-i}^{bi}[h], t_{-i}^{bi}) \succ_i (s_{-i}^{bi}[h'], t_{-i}^{bi})$.

Case 4. Suppose now that h' precedes h. By definition, $s_{-i}^{bi}[h']$ selects the backward induction choices at all information sets weakly following h'. Moreover, by construction of the argument, $s_{-i}^{bi}[h']$ is in $S_{-i}(h)$, so all selected choices weakly following h' lie on the path to h. Hence, we conclude that all opponents' choices between h' and h are backward induction choices. This, however, would imply that $s_{-i}^{bi}[h] = s_{-i}^{bi}[h']$, which is a contradiction to our assumption that $(s_{-i}, t_{-i}) \neq (s_{-i}^{bi}[h], t_{-i}^{bi})$.

Case 5. Suppose finally that h' does not precede nor follow h. That is, h' is parallel to h. Let h^* be the last information set in H_i^* that precedes both h and h'. By definition, $s_{-i}^{bi}[h']$ selects the backward induction choices at all information sets parallel to h'. In particular, $s_{-i}^{bi}[h']$ selects the backward induction choices at all information sets strictly between h^* and h. Moreover, by construction of the argument, $s_{-i}^{bi}[h']$ is in $S_{-i}(h)$, so all selected choices at information sets strictly between h^* and h lie on the path to h. Hence, we conclude that all opponents'

choices between h^* and h are backward induction choices. This implies that $s_{-i}^{bi}[h] = s_{-i}^{bi}[h^*]$. As h' follows h^* we have, by (c) in the statement of the theorem, that $(s_{-i}^{bi}[h^*], t_{-i}^{bi}) \succ_i (s_{-i}^{bi}[h'], t_{-i}^{bi})$, and hence $(s_{-i}^{bi}[h], t_{-i}^{bi}) \succ_i (s_{-i}^{bi}[h'], t_{-i}^{bi})$.

This covers all possible cases. Hence, we see that, indeed, $(s_{-i}^{bi}[h], t_{-i}^{bi}) \succ_i (s_{-i}, t_{-i})$ for all $(s_{-i}, t_{-i}) \in S_{-i}(h) \times T_{-i}$ with $(s_{-i}, t_{-i}) \neq (s_{-i}^{bi}[h], t_{-i}^{bi})$, proving (8). This, in turn, implies (7). Hence, we conclude that the reasoning context of 'common belief in rationality at future and parallel information sets' is characterized by the profile $(\succcurlyeq_i)_{i \in I}$ of plausibility orderings. This completes the proof. ■