Epistemic Foundations for Backward Induction: An Overview

Andrés Perea

Department of Quantitative Economics
Universiteit Maastricht
6200 MD Maastricht, The Netherlands
a.perea@ke.unimaas.nl

Abstract

In this survey we analyze and compare various sufficient epistemic conditions for backward induction that have been proposed in the literature. To this purpose we present a simple epistemic base model for games with perfect information, and express the conditions of the different models in terms of our base model. This will enable us to explictly analyze the differences and similarities between the various sufficient conditions for backward induction.

1 Introduction

Backward induction constitutes one of the oldest concepts in game theory. Its algorithmic definition, which goes back at least to [Ze13], seems so natural at first sight that one might be tempted to argue that every player "should" reason in accordance with backward induction in every game with perfect information. However, on a decision theoretic level the concept is no longer as uncontroversial as it may seem. The problem is that the backward induction algorithm, when applied from a certain decision node on, completely ignores the history that has led to this decision node, as it works from the terminal nodes towards this decision node. At the same time, the beliefs that the player at this decision node has about his opponents' future behavior may well be affected by the history he has observed so far. For instance, a player who observes that an opponent has not chosen in accordance with backward induction in the past may have a valid reason to believe that this same opponent will continue this pattern in the game that lies ahead. However, such belief revision policies are likely to induce choices that contradict backward induction. We therefore need to impose some non-trivial conditions on the players' belief revision policies in order to arrive at backward induction.

During the last decade or so, the game-theoretic literature has provided us with various epistemic models for dynamic games in which sufficient

Johan van Benthem, Dov Gabbay, Benedikt Löwe (eds.). *Interactive Logic. Proceedings of the 7th Augustus de Morgan Workshop, London.* Texts in Logic and Games 1, Amsterdam University Press 2007, pp. 159–193.

epistemic conditions for backward induction have been formulated. The objective of this survey is to discuss these conditions individually, and to explicitly compare the different conditions with each other. The latter task is particularly difficult since the literature exhibits a large variety of epistemic models, each with its own language, assumptions and epistemic operators. Some models are syntactic while others are semantic, and among the semantic models some are based on the notion of states of the world while others use types instead. As to the epistemic operators, some models apply knowledge operators while others use belief operators, and there is also a difference with respect to the "timing" of these operators. Are players entitled to revise their knowledge or belief during the course of the game, and if so, at which instances can they do so? Different models provide different answers to these, and other, questions.

As to overcome these problems we present in Section 2 an epistemic base model, which will be used as a reference model throughout this overview. In Section 3 we then provide for each of the papers to be discussed a brief description of the model, followed by an attempt to formulate its epistemic conditions for backward induction in terms of our base model. By doing so we formulate all sufficient conditions for backward induction in the same base model, which makes it possible to explicitly analyze the differences and similarities between the various conditions.

Finally, a word about the limitations of this paper. In this survey, we restrict attention to epistemic conditions that lead to the *backward induction strategies for all players.* There are alternative models that lead to the *backward induction outcome,* but not necessarily to the backward induction strategy for each player. For instance, [$\text{Ba}_7\text{Si}_1$02] and [$\text{Br}_1\text{Fr}_2\text{Ke}_1$04] provide epistemic models for extensive form rationalizability [$\text{Pe}_0$84, $\text{Ba}_7$97] and iterated maximal elimination of weakly dominated strategies, respectively, which always lead to the backward induction outcome in every generic game with perfect information, but not necessarily to the backward induction strategy profile. We also focus exclusively on sufficient conditions that apply to *all* generic games with perfect information. There are other interesting papers that deal with the logic of backward induction in *specific* classes of games, such as Rosenthals's centipede game [$\text{Ro}_3$81] and the finitely repeated prisoner's dilemma. See, among others, [$\text{Bi}_1$87, $\text{St}_1$96, Au98, $\text{Ra}_0$98, $\text{Br}_3\text{Ra}_0$99, $\text{Ca}_2$00, Pr00]. We shall, however, not discuss these papers here. Even with the limitations outlined above, we do not claim to offer an exhaustive list of epistemic models for backward induction. We do believe, however, that the list of models treated here will give the reader a good impression of the various epistemic conditions for backward induction that exist in the literature.

2 An epistemic base model

2.1 Games with perfect information

A dynamic game is said to be with *perfect information* if every player, at each instance of the game, observes the opponents' moves that have been made until then. Formally, an *extensive form structure $\mathcal{S}$ with perfect information* consists of the following ingredients:

- First, there is a *rooted, directed tree* $T = (X, E)$, where X is a finite set of nodes, and $E \subseteq X \times X$ is a finite set of directed edges. The nodes represent the different situations that may occur during the game, and the edges (x, y) represent moves by players that carry the game from situation x to situation y. The root $x_0 \in X$ marks the beginning of the game. For every two nodes $x, y \in X$, there is at most one path $((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ in E from x to y with $x_1 = x$, $y_n = y$, and $y_k = x_{k+1}$ for every $k \in \{1, \dots, n-1\}$. We say that x precedes y (or, y follows x) if there is a path from x to y. Since x_0 is the root, there is for every $x \in X \backslash \{x_0\}$ a path from x_0 to x. A node $x \in X$ is called a *terminal node* if it is not followed by any other node in X. The set of terminal nodes is denoted Z, and represents the set of possible outcomes for the game.

- There is a finite set I of *players*, and a *move function* $m : X \backslash Z \to I$ which specifies for every non-terminal node x the player $m(x) \in I$ who has to move at x. For every player i, the set of nodes

  $$H_i := \{x \in X \backslash Z \mid m(x) = i\}$$

  is called the set of *information sets*[1] for player i. Since every information set $h_i \in H_i$ corresponds to a single node it is assumed that a player, whenever it is his turn to move, knows exactly which node in the game tree has been reached. That is, the game has *perfect information.* The root x_0 is identified with the information set h_0, and by $H_i^* := H_i \cup \{h_0\}$ we denote the set of player i information sets, together with the beginning of the game. By $H = \bigcup_{i \in I} H_i$ we denote the set of all information sets.

- For every player i and information set $h_i \in H_i$, the set of edges

  $$A(h_i) := \{(h_i, y) \mid y \in X,\ (h_i, y) \in E\}$$

  is called the set of *actions*, or moves, available at h_i. By $A = \bigcup_{h \in H} A(h)$ we denote the set of all actions.

[1] Note that in our restricted setting an information set consists of a single node. We could therefore also have used the term "non-terminal node" instead of "information set". In particular, the collection of all information sets coincides with the set of all non-terminal nodes.

We now turn to the definition of a strategy. Intuitively, a strategy for player i is a plan that describes what player i would do in every possible situation in the game where it is his turn to move. The formal definition of a strategy we shall employ coincides with the concept of a *plan of action*, as discussed in [Ru$_1$91]. The difference with the usual definition is that we require a strategy only to prescribe an action at those information sets that the same strategy does not avoid. Formally, let $\tilde{H}_i \subseteq H_i$ be a collection of player i information sets, not necessarily containing all information sets, and let $s_i : \tilde{H}_i \to A$ be a mapping prescribing at every $h_i \in \tilde{H}_i$ some available action $s_i(h_i) \in A(h_i)$. For a given information set $h \in H$, not necessarily belonging to player i, we say that s_i *avoids* h if on the path

$$((x_1, y_1), \ldots, (x_n, y_n))$$

from h_0 to h there is some node x_k in $\tilde{H}_i$ with $s_i(x_k) \neq (x_k, y_k)$. That is, the prescribed action $s_i(x_k)$ deviates from this path. Such a mapping $s_i : \tilde{H}_i \to A$ is called a *strategy* for player i if $\tilde{H}_i$ is exactly the collection of player i information sets not avoided by s_i. Obviously, every strategy s_i can be obtained by first prescribing an action at all player i information sets, that is, constructing a strategy in the classical sense, and then deleting from its domain those player i information sets that are avoided by it. For a given strategy $s_i \in S_i$, we denote by $H_i(s_i)$ the collection of player i information sets that are not avoided by s_i. Let S_i be the set of player i strategies. For a given information set $h \in H$ and player i, we denote by $S_i(h)$ the set of player i strategies that do not avoid h. Then, it is clear that a profile $(s_i)_{i \in I}$ of strategies reaches an information set h if and only if $s_i \in S_i(h)$ for all players i.

2.2 Preferences, beliefs and types

The basic assumption in our base model is that every player has a *strict*[2] preference relation over the terminal nodes, and holds at each of his information sets a conditional belief about the opponents' strategy choices and preference relations. In particular, we allow for the fact that players may revise their beliefs about the opponents' preferences as the game proceeds. In order to keep our model as "weak" as possible, we assume that this conditional belief can be expressed by a *set* of opponents' strategies and preference relations. This set represents the strategies and preference relations that the player deems *possible* at his information set. We thus do not consider probabilities, and it is therefore sufficient to specify the players' *ordinal* preferences over terminal nodes. Not only does a player hold first-order conditional beliefs about the opponents' choices and preferences,

[2] In the literature, it is not always assumed that players hold strict preferences over terminal nodes. We do so here for the sake of simplicity.

he also holds second-order conditional beliefs about the opponents' possible first-order beliefs at each of his information sets. A second-order belief may thus contain expressions of the form "player i considers it possible at information set h_i that player j considers it possible at information set h_j that player k chooses strategy s_k and has preference relation P_k". Recursively, one may define higher-order conditional beliefs for the players. A possible way to represent such hierarchies of conditional beliefs is by means of the following model.

Definition 2.1 (Epistemic base model)**.** Let $\mathcal{S}$ be an extensive form structure with perfect information. An *epistemic base model* for $\mathcal{S}$ is a tuple

$$\mathcal{M} = (T_i, P_i, B_i)_{i \in I}$$

where

(1) T_i is a set of *types* for player i;

(2) P_i is a function that assigns to every $t_i \in T_i$ some complete, strict and transitive preference relation $P_i(t_i)$ over the terminal nodes;

(3) B_i is a function that assigns to every $t_i \in T_i$ and every information set $h_i \in H_i^*$ some subset $B_i(t_i, h_i) \subseteq \prod_{j \neq i}(S_j(h_i) \times T_j)$.

Here, $B_i(t_i, h_i)$ denotes the set of opponents' strategy-type pairs which t_i deems possible at h_i. We denote by $B_i(t_i, h_i|S_{-i})$ the projection of $B_i(t_i, h_i)$ on $\prod_{j \neq i} S_j(h_i)$. That is, $B_i(t_i, h_i|S_{-i})$ is t_i's belief at h_i about the opponents' strategy choices. For any player $j \neq i$, we denote by $B_{ij}(t_i, h_i)$ the projection of $B_i(t_i, h_i)$ on the set $S_j(h_i) \times T_j$. Hence, $B_{ij}(t_i, h_i)$ is t_i's belief at h_i about player j's strategy-type pair.

From an epistemic base model, conditional beliefs of any order can be *derived*. For instance, type t_i's belief at h_i about player j's choice is given by the projection of $B_{ij}(t_i, h_i)$ on S_j. Let $B_{ij}(t_i, h_i|S_j)$ denote this projection, and let $B_{ij}(t_i, h_i|T_j)$ denote its projection on T_j. Then, type t_i's belief at h_i about player j's belief at h_j about player ℓ's choice is given by

$$\bigcup_{t_j \in B_{ij}(t_i, h_i|T_j)} B_{j\ell}(t_j, h_j|S_\ell).$$

In a similar fashion, higher-order beliefs can be derived.

2.3 Common belief

Let $\mathcal{M} = (T_i, P_i, B_i)_{i \in I}$ be an epistemic base model, and $E \subseteq \bigcup_{j \in I} T_j$ a set of types, or *event*. We say that type t_i *believes in E* at information set $h_i \in H_i^*$ if $B_{ij}(t_i, h_i|T_j) \subseteq E$ for all $j \neq i$. We say that t_i *initially believes* *in* E if t_i believes in E at h_0. Common belief in the event E is defined by the following recursive procedure:

$$\mathrm{B}_i^1(E) = \{t_i \in T_i \mid B_{ij}(t_i, h_i|T_j) \subseteq E \text{ for all } j \neq i \text{ and all } h_i \in H_i^*\}$$

for all $i \in I$, and

$$\mathrm{B}_i^{k+1}(E) = \{t_i \in T_i \mid B_{ij}(t_i, h_i|T_j) \subseteq B_j^k(E) \text{ for all } j \neq i \text{ and all } h_i \in H_i^*\}$$

for all $i \in I$ and all $k \geq 1$.

Definition 2.2 (Common belief)**.** A type $t_i \in T_i$ is said to respect *common belief* in the event E if $t_i \in E$ and $t_i \in \mathrm{B}_i^k(E)$ for all k.

Hence, t_i respects common belief in E if t_i belongs to E, believes throughout the game that opponents' types belong to E, believes throughout the game that opponents believe throughout the game that the other players' types belong to E, and so on. In particular, if we say that t_i respects common belief in E, this implies that t_i itself should belong to E. So, for instance, if we say that t_i respects common belief in the event that types never change their belief about the opponents' preference relations during the game, this implies that t_i itself never changes its belief about the opponents' preference relations. We realize that this is perhaps linguistically not correct, but we do so for the sake of brevity. Otherwise, we should have to write, throughout this paper, that t_i belongs to E, and respects common belief in E.

In most other epistemic models in the literature, the term "common belief" or "common knowledge" refers to the epistemic state of a *group* of players, rather than to the epistemic state of a single player, as we use it. That is, in most other models the expression "there is common belief in E" means that "all players believe that E holds, all players believe that all players believe that E holds, and so on." The reason for me to use an "individualistic" version of common belief is that we want to impose conditions on the beliefs of *one player only*, and see when such individual conditions lead to backward induction reasoning by that player. So, we take a single-player perspective in this paper, and choose the version of common belief accordingly. We realize this is an unusual approach, but it is well-suited for the purposes we have in mind.

Common initial belief in the event E is defined as follows:

$$\mathrm{IB}_i^1(E) = \{t_i \in T_i \mid B_{ij}(t_i, h_0|T_j) \subseteq E \text{ for all } j \neq i\}$$

for all $i \in I$, and

$$\mathrm{IB}_i^{k+1}(E) = \{t_i \in T_i \mid B_{ij}(t_i, h_0|T_j) \subseteq \mathrm{IB}_j^k(E) \text{ for all } j \neq i\}$$

for all $i \in I$ and all $k \geq 1$.

Definition 2.3 (Common initial belief)**.** A type $t_i \in T_i$ is said to respect *common initial belief* in the event E if $t_i \in E$ and $t_i \in \mathrm{IB}_i^k(E)$ for all k.

2.4 Belief in the opponents' rationality

All the epistemic foundations for backward induction to be discussed here make assumptions about the beliefs that players have about the rationality of their opponents. More precisely, all foundations require that players *initially* believe that each opponent chooses rationally at every information set. However, the various foundations differ as to how players would *revise* their beliefs upon observing that their initial belief about the opponents was incorrect. In order to express these different belief revision procedures in terms of our base model, we need the following definitions.

We first define what it means that a strategy is rational for a type at a given information set. For an information set $h_i \in H_i$, a strategy $s_i \in S_i(h_i)$, and an opponents' strategy profile $s_{-i} \in \prod_{j \neq i} S_j(h_i)$, let $z(s_i, s_{-i}|h_i)$ be the terminal node that would be reached from h_i if (s_i, s_{-i}) were to be executed by the players.

Definition 2.4 (Rationality at an information set)**.** A strategy s_i is *rational* for type t_i at information set $h_i \in H_i(s_i)$ if there is no $s_i' \in S_i(h_i)$ such that $P_i(t_i)$ ranks $z(s_i', s_{-i}|h_i)$ strictly over $z(s_i, s_{-i}|h_i)$ for all $s_{-i} \in B_i(t_i, h_i|S_{-i})$.

Hence, s_i is rational for t_i at h_i is there is no other strategy $s_i' \in S_i(h_i)$ that strictly dominates s_i, given the set of opponents' strategies that t_i deems possible at h_i.

We shall now define various restrictions on the beliefs that players have about the opponents' rationality. We need one more definition to this purpose. For a given type $t_i \in T_i$, information set $h_i \in H_i^*$, and some opponent's information set $h \in H \backslash H_i$ following h_i, we say t_i *believes* h *to be reached from* h_i if $B_i(t_i, h_i|S_{-i}) \subseteq S_{-i}(h)$. Here, $S_{-i}(h)$ is a short way to write $\prod_{j \neq i} S_j(h)$.

Definition 2.5 (Belief in the opponents' rationality)**.**

(1) Type t_i believes at information set $h_i \in H_i^*$ that player j chooses rationally at information set $h_j \in H_j$ if for every $(s_j, t_j) \in B_{ij}(t_i, h_i)$ it is true that s_j is rational for t_j at h_j.

(2) Type t_i *initially* believes in rationality at *all* information sets if t_i believes at h_0 that every opponent j chooses rationally at all $h_j \in H_j$.

(3) Type t_i *always* believes in rationality at *all future* information sets if t_i believes at every $h_i \in H_i^*$ that every opponent j chooses rationally at every $h_j \in H_j$ that follows h_i.

(4) Type t_i *always* believes in rationality at *future information sets that are believed to be reached* if t_i believes at every $h_i \in H_i^*$ that every opponent j chooses rationally at all those $h_j \in H_j$ following h_i which t_i believes to be reached from h_i.

(5) Type t_i *always* believes in rationality at *all future and parallel* information sets if t_i believes at every $h_i \in H_i^*$ that every opponent j chooses rationally at every $h_j \in H_j$ that does not precede h_i.

(6) Type t_i *always* believes in rationality at *all* information sets if t_i believes at every $h_i \in H_i^*$ that every opponent j chooses rationally at every $h_j \in H_j$.

Condition (6) is the strongest possible condition that can be imposed, since it requires a player, under all circumstances, to maintain his belief that his opponents have chosen rationally in the past, will choose rationally at any stage in the future, and would have chosen rationally at all foregone (i.e., parallel) information sets. In particular, a player is assumed to interpret every observed past move as being part of an opponent's strategy which is rational at all information sets. In other words, every past move is interpreted as a rational move.

Condition (5) is a weakening of (6). In condition (5), a player need not interpret every observed past move as a rational move, since he is no longer required to believe in the opponents' rationality at past information sets. That is, if a player observes an opponent's move which surprises him, then he may believe that this move was due to a mistake by the opponent. However, condition (5) still requires the player to believe that this same opponent will choose rationally all stages in the future, and would have chosen rationally at all foregone situations. Hence, in condition (5) an observed surprising move by an opponent should not be a reason for dropping the belief in this opponent's rationality at future and foregone situations.

Condition (3) is a weakening of (5), since it no longer requires that a player, after observing a surprising move by an opponent, believes that this opponent would have chosen rationally at foregone situations. However, the player is still assumed to believe that the opponent will, and would, choose rationally at all future situations, no matter whether he deems these future situations possible or not.

Condition (4), in turn, is a weakening of (3). In condition (4) a player, after observing a surprising move by an opponent, need not believe that this opponent would choose rationally at future situations which he does not deem possible. He is only required to believe that the opponent will choose rationally at future situations which he indeed believes could take place.

Condition (2) is a weakening of (3) but not of (4). In condition (2), a player believes, before anything has happened, that every opponent will, and would, choose rationally at all future situations, no matter whether he deems these situations possible or not. However, once the game is under way, the player is allowed to completely drop his belief in the opponents' rationality. Note than in condition (4), a player may initially believe that an opponent would not choose rationally at a certain information set if he initially believes that this information set will not be reached. Therefore, condition (2) is not a weakening of (4). Also, condition (4) is not a weakening of (2), so there is no logical implication betweens conditions (2) and (4).

In light of the definitions we have seen so far, we may thus construct phrases as "type t_i respects common belief in the event that all types initially believe in rationality at all information sets". Some of the epistemic foundations for backward induction, however, use a condition that cannot be expressed in this form, since it relies on a notion that is different from common belief. In order to formalize this condition, we consider the following recursive procedure:

$$\begin{aligned}\mathrm{FBSR}_i^1(h_i) = \{t_i \in T_i \mid\ & t_i \text{ believes at } h_i \text{ that every } j \neq i \text{ chooses} \\ & \text{rationally at all } h_j \text{ that follow } h_i\}\end{aligned}$$

for all $i \in I$ and all $h_i \in H_i^*$, and

$$\begin{aligned}\mathrm{FBSR}_i^{k+1}(h_i) = \{t_i \in T_i \mid\ & B_{ij}(t_i, h_i|T_j) \subseteq \mathrm{FBSR}_j^k(h_j) \text{ for all } j \neq i \\ & \text{and all } h_j \text{ that follow } h_i\}\end{aligned}$$

for all $i \in I$, $h_i \in H_i^*$ and $k \geq 1$.

Definition 2.6 (Forward belief in substantive rationality)**.** A type t_i is said to respect *forward belief in substantive rationality* if $t_i \in \mathrm{FBSR}_i^k(h_i)$ for all k and all $h_i \in H_i^*$.

That is, t_i respects forward belief in substantive rationality if t_i (1) always believes that every opponent is rational at every future information set, (2) always believes that every opponent, at every future information set, believes that every opponent is rational at every future information set, (3) always believes that every opponent, at every future information set, believes that every opponent, at every future information set, believes that every opponent is rational at every future information set, and so on.

The first condition above, namely that t_i always believes that every opponent is rational at every future information set, corresponds exactly to condition (3) of Definition 2.5. However, forward belief in substantive rationality is logically weaker than common belief in condition (3) of Definition 2.5. Consider, namely, a player i information set h_i and a player j information set h_j that precedes h_i. Then, common belief in condition (3) requires that player i believes at h_i that player j believes at h_j that player i chooses rationally at h_i, since h_i follows h_j. On the other hand, forward belief in substantive rationality does not require this, as it only restricts the belief that player i has at h_i about the beliefs that opponents have at information sets *following* h_i, but not preceding h_i. At the same time, it can be verified that forward belief in substantive rationality implies common *initial* belief in condition (3).

We also present a weaker version of forward belief in rationality, which we call forward belief in *material* rationality. Let $H_j(t_i, h_i)$ be the set of those player j information sets h_j following h_i which t_i believes to be reached from h_i. Consider the following recursive procedure:

$$\begin{aligned}\mathrm{FBMR}_i^1(h_i) = \{t_i \in T_i \mid\ & t_i \text{ believes at } h_i \text{ that every } j \neq i \text{ chooses} \\ & \text{rationally at all } h_j \text{ in } H_j(t_i, h_i)\}\end{aligned}$$

for all $i \in I$ and all $h_i \in H_i^*$, and

$$\begin{aligned}\mathrm{FBMR}_i^{k+1}(h_i) = \{t_i \in T_i \mid\ & B_{ij}(t_i, h_i | T_j) \subseteq \mathrm{FBMR}_j^k(h_j) \text{ for all } j \neq \\ & i \text{ and all } h_j \text{ in } H_j(t_i, h_i)\}\end{aligned}$$

for all $i \in I$, $h_i \in H_i^*$ and $k \geq 1$.

Definition 2.7 (Forward belief in material rationality)**.** A type t_i is said to respect *forward belief in material rationality* if $t_i \in \mathrm{FBMR}_i^k(h_i)$ for all k and all $h_i \in H_i^*$.

The crucial difference with forward belief in substantive rationality is thus that a type is only required to believe his opponents to choose rationally at future information sets *which he believes to be reached.* And a type is only required to believe that the opponents' types believe so at future information sets which he believes to be reached, and so on.

The condition in the first step of the recursive procedure, namely that t_i believes at h_i that every opponent j chooses rationally at all h_j in $H_j(t_i, h_i)$, corresponds exactly to condition (4) of Definition 2.5. However, by the same argument as above for forward belief in substantive rationality, it can be verified that forward belief in material rationality is logically weaker than common belief in condition (4) of Definition 2.5. On the other hand, forward belief in material rationality implies common *initial* belief in condition (4).

3 Epistemic foundations for backward induction

In this section we provide an overview of various epistemic foundations that have been offered in the literature for backward induction. A comparison between these foundations is difficult, since the models used by these foundations differ on many aspects.

A first important difference lies in the way the players' beliefs about the opponents are expressed. Some models express the players' beliefs *directly* by means of logical propositions in some formal language. Other models represent the players' beliefs *indirectly* by a set of states of the world, and assign to each state and every player some strategy choice for this player, together with a belief that the player holds at this state about the state of the world. From this model we can derive the higher-order beliefs that players hold about the opponents' choices and beliefs. There are yet some other models that represent the players' beliefs indirectly by means of types, and assign to every type some belief about the other players' choices and types. Similarly to the previous approach, the players' higher-order beliefs can be derived from this model. We refer to these three approaches as the *syntactic model,* the *state-based semantic model* and the *type-based syntactic model.* Note that our base model from the previous section belongs to the last category. This choice is somewhat arbitrary, since we could as well have chosen a syntactic or state-based semantic base model.

Even within the state-based semantic model, the various papers differ on the precise formalization of the beliefs that players have about the state of the world. Similarly, within the type-based model different papers use different belief operators expressing the players' beliefs about the opponents' choices and types.

Finally, some models impose additional conditions on the extensive form structure, such as one information set per player, or the presence of only two players, whereas other papers do not.

In spite of these differences, all foundations have two aspects in common. First, all models provide a theorem, say Theorem A, which gives a sufficient condition for backward induction. Hence, Theorem A states that if player i's belief revision procedure about the other players' choices, preferences and beliefs satisfies some condition **BR**, then his unique optimal choice is his backward induction choice. Secondly, all models guarantee that this sufficient condition **BR** is possible. That is, each paper provides a second result, say Theorem B, which states that for every player i there is some model in which player i's belief revision procedure satisfies condition **BR**. As we shall see, the various foundations differ in the sufficient condition **BR** that is being employed.

In order to explicitly compare the different foundations for backward induction, we attempt to express the various conditions **BR** used by the different models in terms of our base model. By doing so, we express the Theorems A and B used by the various foundations in the following standardized form:

Theorem A. Let $\mathcal{S}$ be an extensive form structure with perfect information, and let $\mathcal{M} = (T_j, P_j, B_j)_{j\in I}$ be an epistemic base model for $\mathcal{S}$. Let

$(\tilde{P}_j)_{j \in I}$ be a profile of strict preference relations over the terminal nodes. If type $t_i \in T_i$ has preference relation $\tilde{P}_i$, and if t_i's conditional belief vector about the opponents' strategy choices and types satisfies condition **BR,** then there is a unique strategy that is rational for t_i at all information sets, namely his backward induction strategy in the game given by $(\tilde{P}_j)_{j \in I}$.

Theorem B. Let $\mathcal{S}$ be an extensive form structure with perfect information, and let i be a player. Then, there is some epistemic base model $\mathcal{M} = (T_j, P_j, B_j)_{j \in I}$ for $\mathcal{S}$ and some type $t_i \in T_i$ such that t_i's conditional belief vector satisfies **BR.**

In the overview that follows, we provide a brief description of every model, identify the condition **BR** that is being used, and explain how this condition may be expressed in terms of our base model. The models are put in alphabetical order.

3.1 Asheim's model

Asheim uses a type-based semantic model, restricted to the case of two players, in which the players' beliefs are modelled by lexicographic probability distributions [As02]. Formally, an Asheim model is given by a tuple

$$\mathcal{M} = (T_i, v_i, \lambda_i)_{i \in I}$$

where T_i is a finite set of types, v_i is a function that assigns to every t_i some von Neumann-Morgenstern utility function $v_i(t_i)$ over the set of terminal nodes, and λ_i is a function that assigns to every type t_i some lexicographic probability system $\lambda_i(t_i)$ on $S_j \times T_j$ with full support on S_j. Such a *lexicographic probability system* $\lambda_i(t_i)$ is given by a vector $(\lambda_i^1(t_i), \ldots, \lambda_i^{K_i(t_i)}(t_i))$ of probability distributions on $S_j \times T_j$. The interpretation is that $\lambda_i^1(t_i), \ldots, \lambda_i^{K_i(t_i)}(t_i)$ represent different degrees of beliefs, and that the kth degree belief $\lambda_i^k(t_i)$ is infinitely more important than the $(k+1)$st degree belief $\lambda_i^{k+1}(t_i)$, without completely discarding the latter. The lexicographic probability system $\lambda_i(t_i)$ induces in a natural way first-order conditional beliefs about player j's choices, as defined in our base model. Namely, for every $h_i \in H_i^*$, let $k_i(t_i, h_i)$ be the first k such that $\lambda_i^k(t_i)$ assigns positive probability to some strategy $s_j \in S_j(h_i)$, and let $\hat{B}_{ij}(t_i, h_i) \subseteq S_j(h_i)$ be the set of strategies in $S_j(h_i)$ to which $\lambda_i^{k_i(t_i,h_i)}(t_i)$ assigns positive probability. Then, t_i induces the conditional belief vector $(\hat{B}_{ij}(t_i, h_i))_{h_i \in H_i^*}$ about player j's strategy choice. For every h_i, let $\hat{T}_{ij}(t_i, h_i) \subseteq T_j$ be the set of types to which $\lambda_i^{k_i(t_i,h_i)}(t_i)$ assigns positive probability. Then, the induced second-order belief of t_i at h_i about player j's belief at h_j about player i's choice is given by the union of the sets $\hat{B}_{ji}(t_j, h_j)$ with $t_j \in \hat{T}_{ij}(t_i, h_i)$.

Similarly, higher-order beliefs about strategy choices can be derived from Asheim's model.

In Asheim's model, a strategy s_i is called rational for type $t_i \in T_i$ at information set h_i if s_i is optimal with respect to the utility function $v_i(t_i)$ and the lexicographic probability system $\lambda_i(t_i|h_i)$, where $\lambda_i(t_i|h_i)$ denotes the conditional of the lexicographic probability system $\lambda_i(t_i)$ on $S_j(h_i) \times T_j$. In particular, if s_i is rational for t_i at h_i then s_i is rational with respect to the preference relation $\hat{P}_i$ and the set-valued belief $\hat{B}_{ij}(t_i, h_i)$, as defined above, where $\hat{P}_i$ is the preference relation on terminal nodes induced by $v_i(t_i)$.

Asheim's sufficient condition for backward induction is based on the notion of *admissible subgame consistency.* A type t_i in an Asheim model is said to be *admissible subgame consistent* with respect to a given profile $(\tilde{v}_j)_{j \in I}$ of utility functions if (1) $v_i(t_i) = \tilde{v}_i$, and (2) for every $h_i \in H_i^*$, the probability distribution $\lambda_i^{k_i(t_i,h_i)}(t_i)$ only assigns positive probability to strategy-type pairs (s_j, t_j) such that s_j is rational for t_j at all $h_j \in H_j$ that follow h_i. In terms of our base model, this condition can be expressed as: (1') $P_i(t_i) = \tilde{P}_i$, and (2') t_i always believes in rationality at all future information sets. In fact, condition (2') is weaker than condition (2) since the notion of rationality in (2') is weaker than the notion of rationality in (2), but condition (2') would have sufficed to prove Asheim's theorem on backward induction.

In Proposition 7, Asheim shows that if a type t_i respects common certain belief in the event that types are admissible subgame consistent with respect to $(\tilde{v}_j)_{j \in I}$, then t_i has a unique strategy that is rational at all information sets, namely his backward induction strategy with respect to $(\tilde{v}_j)_{j \in I}$. Here, "certain belief in an event E" means that type t_i, in each of his probability distributions $\lambda_i^k(t_i)$, only assigns positive probability to types in E. In terms of our base model, this means that the type believes the event E at each of his information sets. In Proposition 8, Asheim shows that common certain belief in admissible subgame consistency is possible. Expressed in terms of our base model, Asheim's sufficient condition for backward induction may thus be written as follows:

Asheim's condition BR: Type t_i respects common belief in the events that (1) types hold preference relations as specified by $(\tilde{P}_j)_{j \in I}$, and (2) types always believe in rationality at all future information sets.

3.2 Asheim & Perea's model

In [AsPe$_2$05], Asheim and Perea propose a type-based semantic model that is very similar to the model from Section 3.1. Attention is restricted to two-player games, and an Asheim-Perea model corresponds to a tuple

$$\mathcal{M} = (T_i, v_i, \lambda_i, \ell_i)_{i \in I},$$

where T_i, v_i and λ_i are as in Asheim's model, and ℓ_i is a function that to every type t_i and event $E \subseteq S_j \times T_j$ assigns some number $\ell_i(t_i, E) \in \{1, \ldots, K_i(t_i)\}$. (Recall that $K_i(t_i)$ denotes the number of probability distributions in $\lambda_i(t_i)$). The interpretation of ℓ_i is that $\ell_i(t_i, E)$ specifies the number of probability distributions in $\lambda_i(t_i)$ that are to be used in order to derive the conditional lexicographic probability system of $\lambda_i(t_i)$ on E. For instance, if player i observes that his information set h_i has been reached, this would correspond to the event $E = S_j(h_i) \times T_j$. In this case, player i would use the first $\ell_i(t_i, E)$ probability distributions in $\lambda_i(t_i)$ in order to form his conditional belief about j upon observing that h_i has been reached. Two extreme cases are $\ell_i(t_i, E) = k_i(t_i, h_i)$, where player i would only use the first probability distribution in $\lambda_i(t_i)$ that assigns positive probability to some player j strategy in $S_j(h_i)$, and $\ell_i(t_i, E) = K_i(t_i)$, where player i would use the full lexicographic probability system $\lambda_i(t_i)$ to form his conditional belief upon observing that h_i is reached. Recall that $k_i(t_i, h_i)$ is the first k such that $\lambda_i^k(t_i)$ assigns positive probability to some strategy $s_j \in S_j(h_i)$.

The sufficient condition for backward induction is based on the event that *types induce for every opponent's type a sequentially rational behavior strategy.* Consider a type t_i, and let $T_j^{t_i}$ be the set of types to which the lexicographic probability system $\lambda_i(t_i)$ assigns positive probability (in some of its probability distributions). Asheim and Perea assume that for every $t_j \in T_j^{t_i}$ and every $s_j \in S_j$, the lexicographic probability system $\lambda_i(t_i)$ assigns positive probability to (s_j, t_j). For every information set $h_j \in H_j$ and action $a \in A(h_j)$, let $S_j(h_j, a)$ be the set of strategies in $S_j(h_j)$ that select action a at h_j. Define for every type $t_j \in T_j^{t_i}$, $h_j \in H_j$ and $a \in A(h_j)$

$$\sigma_j^{t_i|t_j}(h_j)(a) := \frac{\lambda_i^k(t_i)(S_j(h_j, a) \times \{t_j\})}{\lambda_i^k(t_i)(S_j(h_j) \times \{t_j\})},$$

where k is the first number such that $\lambda_i^k(t_i)(S_j(h_j) \times \{t_j\}) > 0$. The vector

$$\sigma_j^{t_i|t_j} = (\sigma_j^{t_i|t_j}(h_j)(a))_{h_j \in H_j, a \in A(h_j)}$$

is called the *behavior strategy induced by t_i for t_j.* My interpretation of $\sigma_j^{t_i|t_j}(h_j)(a)$ is that it describes t_i's conditional belief about t_j's action choices at future and parallel information sets. Let me explain. Consider an information set h_i for player i and an information set h_j for player j which either follows h_i or is parallel to h_i. Then, my interpretation of $\sigma_j^{t_i|t_j}(h_j)(a)$ is that type t_i believes at h_i that type t_j at information h_j chooses action a with probability $\sigma_j^{t_i|t_j}(h_j)(a)$. Namely, the information that the game has reached h_i does not give type t_i additional information about the action choice of t_j at h_j, and hence $\sigma_j^{t_i|t_j}(h_j)$ provides an intuitive candidate for the conditional belief of t_i at h_i about t_j's behavior at h_j.

However, if h_j precedes h_i, then $\sigma_j^{t_i|t_j}(h_j)(a)$ does not necessarily describe t_i's belief at h_i about t_j's action choice at h_j. In this case, there is namely a unique action a^* at h_j that leads to h_i, whereas $\sigma_j^{t_i|t_j}(h_j)(a^*)$ may be less than one (in fact, may be zero). On the other hand, it should be clear that t_i should believe (with probability 1) at h_i that t_j has chosen a^* at h_j, since it is the only action at h_j that leads to h_i. Hence, in this case $\sigma_j^{t_i|t_j}(h_j)(a)$ cannot describe t_i's belief at h_i about t_j's choice at h_j.

For every information set $h_j \in H_j$, let $\sigma_j^{t_i|t_j}|_{h_j}$ be the behavioral strategy that assigns probability one to all player j actions preceding h_j, and coincides with $\sigma_j^{t_i|t_j}$ otherwise. The induced behavior strategy $\sigma_j^{t_i|t_j}$ is said to be *sequentially rational for* t_j if at every information set $h_j \in H_j$, the behavior strategy $\sigma_j^{t_i|t_j}|_{h_j}$ only assigns positive probability to strategies in $S_j(h_j)$ that are rational for t_j at h_j (in the sense of Asheim's model above). Type t_i is said to induce for every opponent's type a sequentially rational behavior strategy if for every $t_j \in T_j^{t_i}$ it is true that $\sigma_j^{t_i|t_j}$ is sequentially rational for t_j. As we have seen above, $\sigma_j^{t_i|t_j}$ represents for every $h_i \in H_i^*$ type t_i's conditional belief at h_i about player j's behavior at future and parallel information sets. The requirement that $\sigma_j^{t_i|t_j}$ always be sequentially rational for t_j thus means that t_i always believes in rationality at all future and parallel information sets.

In Proposition 11, Asheim and Perea show that if a type t_i respects common certain belief in the events that (1) types have utility functions as specified by $(\tilde{v}_j)_{j\in I}$, and (2) types induce for every opponent's type a sequentially rational behavior strategy, then t_i has a unique strategy that is rational at all information sets, namely his backward induction strategy with respect to $(\tilde{v}_j)_{j\in I}$. The existence of such types follows from their Proposition 4 and the existence of a sequential equilibrium. In terms of our base model, Asheim and Perea's sufficient condition may thus be stated as follows:

Asheim & Perea's condition BR: Type t_i respects common belief in the events that (1) types hold preference relations as specified by $(\tilde{P}_j)_{i\in I}$, and (2) types always believe in rationality at all future and parallel information sets.

3.3 Aumann's model

Aumann proposes a state-based semantic model for extensive form structures with perfect information [Au95]. An Aumann model is a tuple

$$\mathcal{M} = (\Omega, (B_i, f_i, v_i)_{i\in I})$$

where Ω represents the set of states of the world, B_i is a function that assigns to every state $\omega \in \Omega$ some subset $B_i(\omega)$ of states, f_i is a function that assigns to every state ω some strategy $f_i(\omega) \in S_i$, and v_i is a function that assigns to every ω some von Neumann-Morgenstern utility function $v_i(\omega)$ on the set of terminal nodes. The functions B_i must have the property that $\omega \in B_i(\omega)$ for all ω, and for all $\omega, \omega' \in \Omega$ it must hold that $B_i(\omega)$ and $B_i(\omega')$ are either identical, or have an empty intersection. Hence, the set $\{B_i(\omega)|\omega \in \Omega\}$ is a partition of Ω. The interpretation is that at state ω, player i believes that the true state is in $B_i(\omega)$. (In fact, Aumann uses the term "knows" rather than "believes"). The functions f_i and v_i must be measurable with respect to B_i, meaning that $f_i(\omega') = f_i(\omega)$ whenever $\omega' \in B_i(\omega)$, and similarly for v_i. The reason is that player i cannot distinguish between states ω and ω', and hence his choice and preferences must be the same at both states.

It is problematic, however, to formally translate this model into conditional beliefs of our base model. Consider, for instance, a game with three players, in which players 1, 2 and 3 sequentially choose between Stay and Leave, and where Leave terminates the game. Consider a state ω where $f_1(\omega) = \mathsf{Leave}$ and $B_2(\omega) = \{\omega\}$. Then, at player 2's information set, player 2 must conclude that the state cannot be ω, but must be some state ω' with $f_1(\omega') = \mathsf{Stay}$. However, there may be many such states ω', and hence it is not clear how player 2 should revise his belief about the state at his information set. Since his revised belief about the state will determine his revised belief about player 3's choice, it is not clear how to explicitly define player 2's revised belief about player 3's choice from Aumann's model.

At the same time, Aumann's Theorem A provides sufficient conditions for backward induction, and hence Aumann's model must at least implicitly impose some restrictions on the players' belief revision procedures, since otherwise backward induction could not be established. The main task in this subsection will be to identify these implicit assumptions about the belief revision procedures, and incorporate these as explicit restrictions in our base model. Since such an identification is a rather subjective procedure, this approach will eventually lead to a model which is a subjective interpretation of Aumann's model.

The model as proposed by Aumann is essentially a static model, since for every state ω and every player i, his belief $B_i(\omega)$ is only defined at a single moment in time. My interpretation of these static beliefs is that players, upon observing that one of their information sets has been reached, do not revise more than "strictly necessary". In fact, the only beliefs that *must* be revised by player i when finding out that his information set h_i has been reached are, possibly, his beliefs about the opponents' choices at information sets preceding h_i. That is, if player 2 in the example above finds out that player 1 has chosen Stay, then this should not be a reason to change his belief about player 3's choice. Even stronger, we interpret Aumann's static beliefs in this game as beliefs in which player 2 *only* changes his belief about player 1's choice, while maintaining all his other beliefs, including his beliefs about the opponents' beliefs. Hence, if we interpret Aumann's model in this way, and express it accordingly in terms of our base model, then every type is supposed to never revise his belief about the opponents' choices and the opponents' beliefs at future and parallel information sets. A type t_i, when arriving at some information set h_i, may only revise his belief about the opponents' *choices* at information sets that precede h_i (but not about their types). For further reference, we call this condition the "no substantial belief revision condition".

The sufficient condition for backward induction presented by Aumann is *common knowledge of rationality.* Let ω be a state, i a player and h_i an information set controlled by i. At state ω, player i is said to be rational at information set h_i if there is no $s_i \in S_i$ such that for every $\omega' \in B_i(\omega)$ it holds that

$$v_i(\omega)(z(s_i,(f_j(\omega'))_{j\neq i}|h_i)) > v_i(\omega)(z(f_i(\omega),(f_j(\omega'))_{j\neq i}|h_i)),$$

where $z(s_i,(f_j(\omega'))_{j\neq i}|h_i)$ is the terminal node that is reached if the game would start at h_i, and the players would choose in accordance with $(s_i,(f_j(\omega'))_{j\neq i})$. In terms of our base model, this means that strategy $f_i(\omega)$ is rational for player i at h_i with respect to the utility function $v_i(\omega)$ and his first-order belief $\{(f_j(\omega'))_{j\neq i} \mid \omega' \in B_i(\omega)\}$ about the opponents' choices after h_i. Let Ω^{rat} be the set of states ω such that at ω all players are rational at each of their information sets.

Common knowledge of rationality can now be defined by the following recursive procedure:

$$\begin{aligned} \text{CKR}^1 &= \Omega^{\text{rat}}; \\ \text{CKR}^{k+1} &= \{\omega \in \Omega \mid B_i(\omega) \subseteq \text{CKR}^k \text{ for all players } i\} \end{aligned}$$

for $k \geq 1$. Then, common knowledge of rationality is said to hold at ω if $\omega \in \text{CKR}^k$ for all k. In Theorem A, Aumann proves that for every profile $(\tilde{v}_j)_{j\in I}$ of utility functions, for every state ω at which common knowledge of $(\tilde{v}_j)_{j\in I}$ and common knowledge of rationality hold, and for every player i, the strategy $f_i(\omega)$ is the backward induction strategy for player i with respect to $(\tilde{v}_j)_{j\in I}$. In Theorem B, Aumann proves that there is an Aumann model and a state ω at which common knowledge of $(\tilde{v}_j)_{j\in I}$ and common knowledge of rationality hold.

In terms of our base model, common knowledge of rationality implies common initial belief in rationality at all information sets. By the latter we mean that a type (1) initially believes that all players choose rationally at all information sets, (2) initially believes that every type initially believes that all players choose rationally at all information sets, and so on. Together with the "no substantial belief revision condition" above, this would imply that a type *always* believes that types initially believe that all players choose rationally at all information sets, and that a type always believes that types always believe that types initially believe that players choose rationally at all information sets, and so on. Hence, a possible interpretation of Aumann's condition of common knowledge of rationality, together with the "no substantial belief revision condition", in our base model would be: common belief in the event that players initially believe in rationality at all information sets. Similarly, common knowledge of $(\tilde{v}_j)_{j\in I}$, together with the "no substantial belief revision condition", could be interpreted as common belief in the event that types have preferences according to $(\tilde{P}_j)_{j\in I}$, where $\tilde{P}_j$ is the preference relation that corresponds to $\tilde{v}_j$. That is, Aumann's sufficient conditions for backward induction could be interpreted as follows in terms of our base model:

Aumann's condition BR: Type t_i respects common belief in the events that (1) types hold preferences as specified by $(\tilde{P}_j)_{j\in I}$, (2) types initially believe in rationality at all information sets, and (3) types never revise their beliefs about the opponents' choices and beliefs at future and parallel information sets.

In [Cl$_0$03], Clausing basically provides a reformulation of Aumann's model and definitions in a syntactic framework. Clausing's sufficient condition for backward induction is a little weaker than Aumann's, since Clausing only requires "true $(k-1)$st level belief" in rationality at all information sets, where k is the maximal length of a path in the game tree, which is weaker than common knowledge of rationality as defined by Aumann. Quesada proves, in [Qu03, Propositions 3.3 & 3.4] that Aumann's backward induction theorem can also be shown without imposing that $\omega \in B_i(\omega)$ for all ω, and without imposing that for all $\omega, \omega' \in \Omega$ it must hold that $B_i(\omega)$ and $B_i(\omega')$ are either identical, or have an empty intersection. That is, Quesada no longer assumes a partition structure, nor does he require that what one believes must be true. The only substantial conditions that [Qu03] imposes on the belief operator is that $f_i(\omega') = f_i(\omega)$ whenever $\omega' \in B_i(\omega)$, and similarly for v_i. Hence, a player must be aware of his choice and his utility function.

However, since the models by Clausing and Quesada [Cl$_0$03, Qu03] are identical in spirit to Aumann's, we omit a formal discussion of these models in this overview.

3.4 Balkenborg & Winter's model

In [Ba$_3$Wi$_2$97], Balkenborg and Winter present a state-based semantic model that is almost identical to Aumann's model, so we do not repeat it here.

The only difference is that Balkenborg and Winter restrict attention to extensive form structures in which every player controls only one information set. However, the sufficient conditions given for backward induction are different from Aumann's conditions, as they are based on the notion of *forward knowledge of rationality* rather than common knowledge of rationality.

For every player i, let h_i be the unique information set controlled by player i. The definition of player i being rational at h_i is the same as in Aumann's model. Let Ω_i^{rat} be the set of states ω such that at ω, player i is rational at h_i. We say that player j comes after player i if h_j comes after h_i. Forward knowledge of rationality can now be defined by the following recursive procedure. For every player i define:

$$\begin{array}{lcl} \text{FKR}_i^1 & = & \Omega_i^{\text{rat}}; \\ \text{FKR}_i^{k+1} & = & \{\omega \in \Omega \mid B_i(\omega) \subseteq \text{FKR}_j^k \text{ for all } j \text{ that come after } i\}, \end{array}$$

for every $k \geq 1$. Then, forward knowledge of rationality is said to hold at state ω if $\omega \in \text{FKR}_i^k$ for all i and all k. That is, player i believes that every player after him will choose rationally, believes that every player after him believes that every player after him will choose rationally, and so on. So, it corresponds to our notion of forward belief in substantive rationality in Definition 2.6.

In Theorem 2.1, Balkenborg and Winter prove that for every profile $(\tilde{v}_j)_{j \in I}$ of utility functions, for every state ω at which common knowledge of $(\tilde{v}_j)_{j \in I}$ and forward knowledge of rationality hold, and for every player i, the strategy $f_i(\omega)$ is the backward induction strategy for player i with respect to $(\tilde{v}_j)_{j \in I}$. Balkenborg and Winter's sufficient condition for backward induction may thus be phrased as follows in terms of our base model:

Balkenborg & Winter's condition BR: Type t_i (1) respects common belief in the event that types hold preferences as specified by $(\tilde{P}_j)_{j \in I}$, (2) respects forward belief in substantive rationality, and (3) respects common belief in the event that types never revise their beliefs about the opponents' choices and beliefs at future and parallel information sets.

Quesada proves in [Qu03, Proposition 3.1] that Balkenborg and Winter's sufficient condition for backward induction would still be sufficient if one weakens the conditions on the knowledge operators as explained at the end of the previous subsection.

3.5 Clausing's model

Clausing presents a syntactic model for games with perfect information [Cl$_0$04]. For our purposes here it is not necessary to discuss the complete formalism of Clausing's model, and therefore we restrict ourselves to presenting only the key ingredients. A major technical difference between our description here and Clausing's original model is that we shall employ "statements" instead of logical propositions. The reader is referred to the original paper for the syntactic formalism employed by Clausing. For our restricted purposes here, a Clausing model may be described as a tuple

$$\mathcal{M} = (L, (\hat{B}_i, v_i)_{i \in I})$$

where L is a language, or set of statements, $\hat{B}_i$ is a function that assigns to every statement $f \in L$ some subset $\hat{B}_i(f) \subseteq L$ of statements, and v_i is a utility function for player i on the set of terminal nodes. By "$g \in \hat{B}_i(f)$" we mean the statement that "player i believes statement g upon learning that f holds". It is assumed that L contains all statements of the form "player i chooses strategy s_i", and that it is closed under the operations $\neg$ (not), $\wedge$ (and) and $\hat{B}_i$. By the latter, we mean that if f and g are statements in L, then so are the statements "$\neg f$", "$f \wedge g$" and "$g \in \hat{B}_i(f)$".

Clausing's sufficient condition for backward induction is *forward belief from the root to all information sets h in rationality at h*. We say that strategy s_i is rational for player i at information set h_i if there is no other strategy $s_i' \in S_i(h_i)$ such that player i would believe, upon learning that h_i has been reached, that s_i' would lead to a higher utility than s_i. Formally, there should be no $s_i' \in S_i(h_i)$ and no statement $f \in L$ about the opponents' strategy choices such that (1) player i believes f upon learning that all opponents j have chosen a strategy in $S_j(h_i)$, and (2) for every opponents' strategy profile s_{-i} compatible with f it would be true that $v_i(z(s_i', s_{-i}|h_i)) > v_i(z(s_i, s_{-i}|h_i))$. Player i is said to believe at h_i that player j is rational at h_j if, upon learning that h_i has been reached, player i believes the statement "player j chooses a strategy that is rational for j at h_j". Forward belief from the root to all information sets h in rationality at h can now be defined by the following sequence of statements:

$\mathrm{FB}_i^1(h_i) =$ "player i believes, upon learning that h_i has been reached, that every opponent j will be rational at all h_j that follow h_i"

for all players i and all $h_i \in H_i^*$, and

$\mathrm{FB}_i^{k+1}(h_i) =$ "player i believes, upon learning that h_i has been reached, the statement $\mathrm{FB}_j^k(h_j)$ for all opponents j and all h_j that follow h_i"

for all players i, $h_i \in H_i^*$ and $k \geq 1$. Player i is said to respect forward belief from the root to all information sets h in rationality at h if for every h_i, player i believes, upon learning that h_i has been reached, the statements $\mathrm{FB}_j^k(h_j)$ for all k, all opponents j and all $h_j \in H_j$ that follow h_i. In Proposition 2, Clausing shows that this condition implies backward induction, whereas his Proposition 3 demonstrates that this condition is possible. In terms of our base model, Clausing's condition clearly corresponds to forward belief in substantive rationality.

Clausing's condition BR: Type t_i (1) respects common belief in the event that types hold preferences as specified by $(\tilde{P}_j)_{j\in I}$, and (2) respects forward belief in substantive rationality.

3.6 Feinberg's model

Feinberg provides a syntactic model for dynamic games which is similar to Clausing's model [Fe$_1$05]. Since a full treatment of Feinberg's model would take us too far afield, we present a highly condensed version of his model here, which will serve for our restricted purposes. As with the discussion of Clausing's model, we refer to the original paper for the syntactic formalism. For our purposes here, a Feinberg model may be described as a tuple

$$\mathcal{M} = (L, (C_i, v_i)_{i\in I})$$

where L is a language, or set of statements, C_i is a function that selects for every information set $h_i \in H_i^*$ a set $C_i(h_i) \subseteq L$ of statements, and v_i is a utility function for player i on the set of terminal nodes. The interpretation of $f \in C_i(h_i)$ is that player i is *confident* of statement f at information set h_i. The language L must contain all statements of the form "player i chooses strategy s_i", and must be closed under the application of the operators $\neg$ (not), $\wedge$ (and) and $C_i(h_i)$. By the latter we mean that, if f is a statement in L, then the statement "$f \in C_i(h_i)$" must also be in L.

Feinberg characterizes the *confidence operator* by means of a list of axioms, which largely coincides with the list of classic axioms for a knowledge operator. The single, but crucial, difference is that a player may be confident of a statement that is objectively wrong, whereas this is not possible in the case of a knowledge operator. However, in Feinberg's model a player must always be confident that he is right, that is, a player must be confident that all statements he is confident of are true.

Feinberg presents two different sufficient conditions for backward induction, namely *common confidence of hypothetical rationality* and *iterated future confidence of rationality.* Strategy s_i is said to be rational for player i at h_i if there is no other strategy $s_i' \in S_i(h_i)$ such that player i would be confident at h_i that s_i' would lead to a higher utility than s_i. By the latter, we mean that there should be no $s_i' \in S_i(h_i)$, and no statement f about the opponents' strategy choices, such that (1) i is confident of f at h_i, and (2) for every opponents' strategy profile s_{-i} compatible with f it would hold that $v_i(z(s_i', s_{-i})|h_i) > v_i(z(s_i, s_{-i})|h_i)$. We say that i is confident at h_i that j is rational at h_j if the statement "player j chooses a strategy that is rational for j at h_j" belongs to $C_i(h_i)$. Common confidence in hypothetical rationality can now be defined recursively by the following sequence of statements:

$$\text{CCHR}^1 = \text{“every player } i \text{ is confident at every } h_i \text{ that every opponent } j \text{ will be rational at every } h_j \text{ not preceding } h_i\text{”}$$

and, for every $k \geq 1$,

$$\text{CCHR}^{k+1} = \text{“every player } i \text{ is confident at every } h_i \text{ of CCHR}^k\text{”}.$$

Player i is said to respect common confidence in hypothetical rationality if, for every h_i and every k, player i is confident at h_i of CCHR^k. In Proposition 10, Feinberg shows that this condition is possible, and implies backward induction. In terms of our base model, this condition corresponds exactly to our definition of common belief in the event that types always believe in rationality at all future and parallel information sets.

Feinberg's first condition BR: Type t_i respects common belief in the events that (1) types hold preference relations as specified by $(\tilde{P}_j)_{j\in I}$, and (2) types always believe in rationality at all future and parallel information sets.

Iterated future confidence of rationality can be defined by means of the following sequence of statements:

$$\text{IFCR}_i^1(h_i) = \text{“player } i \text{ is confident at } h_i \text{ that all opponents } j \text{ will be rational at all } h_j \text{ that follow } h_i\text{”}$$

for all $i \in I$ and all $h_i \in H_i^*$, and

$$\text{IFCR}_i^{k+1}(h_i) = \text{“player } i \text{ is confident at } h_i \text{ of IFCR}_j^k(h_j) \text{ for all opponents } j \text{ and all } h_j \text{ that follow } h_i\text{”}$$

for all $i \in I$, $h_i \in H_i^*$ and $k \geq 1$. Player i is said to respect iterated future confidence of rationality if, for every k, every h_i, every opponent j, and every h_j following h_i, player i is confident at h_i of $\text{IFCR}_j^k(h_j)$. Feinberg shows in his Proposition 11 that this condition is possible and leads to backward induction. In terms of our base model, this condition corresponds to our definition of forward belief in substantive rationality.

Feinberg's second condition BR: Type t_i (1) respects common belief in the event that types hold preferences as specified by $(\tilde{P}_j)_{j\in I}$, and (2) respects forward belief in substantive rationality.

3.7 Perea's model

Perea proposes a type-based semantic model that is very similar to our base model [Pe$_2$05]. The difference is that in [Pe$_2$05], the players' initial and revised beliefs are assumed to be point-beliefs, that is, contain exactly one strategy-type pair for each opponent. Moreover, in [Pe$_2$05] the model is assumed to be *complete* which will be defined below. An important difference between Perea's model and the other models discussed here is that Perea's model explicitly allows for the possibility that players revise their belief about the opponents' preference relations over terminal nodes as the game proceeds. A Perea model is a tuple

$$\mathcal{M} = (T_i, P_i, \hat{B}_i)_{i \in I}$$

where T_i is player i's set of types, P_i assigns to every type $t_i \in T_i$ a strict preference relation $P_i(t_i)$ over the terminal nodes, $\hat{B}_i$ assigns to every type $t_i \in T_i$ and every information set $h_i \in H_i^*$ a belief $\hat{B}_i(t_i, h_i) \subseteq \prod_{j \neq i}(S_j(h_i) \times T_j)$ consisting of exactly one point, and the model $\mathcal{M}$ is *complete.* The assumption that the belief $\hat{B}_i(t_i, h_i)$ consists of exactly one point means that t_i, at every information set h_i, is supposed to consider only one strategy-type pair (s_j, t_j) possible for every opponent j. However, this point-belief may change as the game proceeds. By a complete model, we mean that for every player i, every strict preference relation $\hat{P}_i$ and every belief vector $\tilde{B}_i = (\tilde{B}_i(h_i))_{h_i \in H_i^*}$ consisting of conditional point-beliefs $\tilde{B}_i(h_i)$ as described above, there is some type $t_i \in T_i$ with $P_i(t_i) = \hat{P}_i$ and $\hat{B}_i(t_i, h_i) = \tilde{B}_i(h_i)$ for all h_i. Since types may revise their belief about the opponents' types, and different types may have different preference relations over terminal nodes, Perea's model allows types to revise their belief about the opponents' preference relations over terminal nodes.

Perea's sufficient condition for backward induction is common belief in the events that (1) players initially believe in $(\tilde{P}_i)_{i \in I}$, (2) players initially believe in rationality at all information sets, and (3) the players' belief revision procedures satisfy some form of minimal belief revision. The crucial difference with the other models discussed here is that condition (1) allows players to revise their belief about the opponents' preference relations as the game proceeds. On the other hand, the conditions (2) and (3) as they have been defined in [Pe$_2$05] can be shown to imply that players should *always* believe that every opponent chooses rationally at *all information sets;* a condition that cannot be realized in general if players do not revise their beliefs about the opponents' preference relations.

A type t_i is said to initially believe in $(\tilde{P}_j)_{j \in I}$ if for every opponent j, the initial belief $\hat{B}_i(t_i, h_0)$ about player j consists of a strategy-type pair (s_j, t_j) where $P_j(t_j) = \tilde{P}_j$. In order to formalize condition (3), we need the definition of an elementary statement. A first-order elementary statement about player i is a statement of the form "player i has a certain preference relation" or "player i believes at h_i that opponent j chooses a certain strategy". Recursively, one can define, for every $k \geq 2$, a kth order elementary statement about player i as a statement of the form "player i believes at h_i that φ" where φ is a $(k-1)$st order elementary statement. An elementary statement about player i is then an elementary statement about player i of some order k. Now, let $h_i \in H_i \backslash h_0$, and let h_i' be the information set in H_i^* that precedes h_i and for which no other player i information set is between h_i' and h_i. For every opponent j, let (s_j', t_j') be the strategy-type pair in $\hat{B}_i(t_i, h_i')$, and let (s_j, t_j) be the strategy-type pair in $\hat{B}_i(t_i, h_i)$. Type t_i is said to satisfy minimal belief revision at h_i if for every opponent j the strategy-type pair (s_j, t_j) is such that (1) s_j is rational for t_j at all information sets, (2) there is no other strategy-type pair (s_j'', t_j'') in $S_j(h_i) \times T_j$ satisfying (1) such that t_j'' and t_j' disagree on fewer elementary statements about player j than t_j and t_j' do, and (3) there is no other strategy-type pair (s_j'', t_j'') in $S_j(h_i) \times T_j$ satisfying (1) and (2) such $P_j(t_j'')$ and $P_j(t_j')$ disagree on fewer pairwise rankings of terminal nodes than $P_j(t_j)$ and $P_j(t_j')$ do. It can be shown that this notion of minimal belief revision, together with the condition that players initially believe in rationality at all information sets, imply that a type always believes that his opponents choose rationally at *all* information sets. For the definition of minimal belief revision it is very important that the model $\mathcal{M}$ is assumed to complete. [Pe$_2$05, Theorem 5.1] shows that there is a Perea model which satisfies the sufficient condition listed above. Theorem 5.2 in that paper demonstrates that this sufficient condition leads to backward induction. As such, Perea's sufficient condition for backward induction can be stated as follows in terms of our base model:

Perea's condition BR: Type t_i respects common belief in the events that (1) types hold point-beliefs, (2) types initially believe in $(\tilde{P}_j)_{j \in I}$, (3) types always believe in rationality at all information sets, and (4) types satisfy minimal belief revision.

3.8 Quesada's model

Quesada presents a model for games with perfect information which is neither semantic nor syntactic [Qu02]. The key ingredient is to model the players' uncertainty by means of *Bonanno belief systems* [Bo$_0$92]. A Bonnano belief system is a profile $\beta = (\beta_i)_{i \in I}$, where β_i is a belief vector that assigns to every information set h (not necessarily controlled by player i) some terminal node $\beta_i(h)$ which follows h. The interpretation is that player i, upon learning that the game has reached information set h, believes that he and his opponents will act in such a way that terminal node $\beta_i(h)$ will be reached. A Quesada model is a pair

$$\mathcal{M} = (\mathcal{B}, (v_i)_{i \in I})$$

where $\mathcal{B}$ is a set of Bonnano-belief systems, and v_i is a utility function for player i over the terminal nodes. Quesada's sufficient condition for backward induction states that every belief system in $\mathcal{B}$ should be *rational*, and that every belief system in $\mathcal{B}$ should be *justifiable* by other belief systems in $\mathcal{B}$. Formally, a belief system $\beta = (\beta_i)_{i \in I}$ is said to be rational if for every player i and every information set $h_i \in H_i$ it holds that $v_i(\beta_i(h_i)) \geq v_i(\beta_i((h_i, a)))$ for every action $a \in A(h_i)$, where (h_i, a) denotes the information set that immediately follows action a at h_i. We say that belief system $\beta = (\beta_i)_{i \in I}$ in $\mathcal{B}$ is justifiable by other belief systems in $\mathcal{B}$ if for every player i, every $h_i \in H_i$, every opponent j, and every $h_j \in H_j$ between h_i and the terminal node $\beta_i(h_i)$ there is some belief system $\beta' = (\beta'_i)_{i \in I}$ in $\mathcal{B}$ such that $\beta'_j(h_j) = \beta_i(h_i)$. A belief system $\beta = (\beta_i)_{i \in I}$ is called the backward induction belief system if for every player i and every information set h, $\beta_i(h)$ is the terminal node which is reached by applying the backward induction procedure (with respect to $(v_i)_{i \in I}$) from h onwards. In Proposition 1, Quesada shows that there is one, and only one, set $\mathcal{B}$ which satisfies the two conditions above, namely the set containing only the backward induction belief system.

We shall now attempt to express these conditions in terms of our base model. Take a set $\mathcal{B}$ of belief systems such that every belief system in $\mathcal{B}$ is justifiable by other belief systems in $\mathcal{B}$ (and thus satisfies Quesada's second condition above). Then, every belief system β_i in $\mathcal{B}$ induces, for every h_i, a point-belief about the opponents' strategy choices as follows: For every h_i there is some opponents' strategy profile $s_{-i}(\beta_i, h_i) \in \prod_{j \neq i} S_j(h_i)$ such that, for every action $a \in A(h_i)$, the action a followed by $s_{-i}(\beta_i, h_i)$ leads to the terminal node $\beta_i(h_i, a)$. Hence, $s_{-i}(\beta_i, h_i)$ may be interpreted as β_i's conditional point-belief at h_i about the opponents' strategy choices. (Note that this belief need not be unique, as β_i does not restrict player i's beliefs at h_i about opponents' choices at parallel information sets). The belief vector β_i also induces, for every h_i, a conditional point-belief about the opponents' belief vectors β'_j in $\mathcal{B}$. Consider, namely, an information set $h_i \in H_i$, some opponent j and an information set h_j between h_i and the terminal node $\beta_i(h_i)$ such that there is no further player j information set between h_i and h_j. Since $\mathcal{B}$ satisfies Quesada's justifiability condition, there is some player j belief vector $\beta_j(\beta_i, h_i)$ in $\mathcal{B}$ such that $\beta_j(\beta_i, h_i)(h_j) = \beta_i(h_i)$. (Again, this choice need not be unique). This belief vector $\beta_j(\beta_i, h_i)$ may then serve as β_i's conditional point-belief at h_i about player j's belief vector. Summarizing, every belief vector β_i induces, at every h_i, a conditional point-belief about the opponents' strategy choices and the opponents' belief vectors.

Now, if we interpret every belief vector β_i in $\mathcal{B}$ as a type $t_i(\beta_i)$ in our base model, then, by the insights above, every type $t_i(\beta_i)$ induces, at every h_i, a conditional point-belief about the opponents' strategy choices and types $t_j(\beta_j)$. Hence, similarly to Perea's model, Quesada's model can be expressed in terms of our base model by imposing common belief in the event that types hold point-beliefs. Let $T_i(\mathcal{B})$ denote the set of all such types

$t_i(\beta_i)$ induced by some belief vector β_i in $\mathcal{B}$. A combination of Quesada's rationality condition and justifiability condition implies that, whenever β_i in $\mathcal{B}$ believes at h_i that player j chooses action a at some h_j between h_i and $\beta_i(h_i)$ (with no player j information set between h_i and h_j), then there is some rational belief vector $\beta_j(\beta_i, h_i)$ in $\mathcal{B}$ such that $\beta_j(\beta_i, h_i)(h_j) = \beta_i(h_i)$. In particular, action a must be part of the rational belief vector $\beta_j(\beta_i, h_i)$, and hence action a must be optimal with respect to $\beta_j(\beta_i, h_i)$. In terms of our base model, this means that, whenever type $t_i(\beta_i)$ believes at h_i that information set h_j will be reached in the future, and believes at h_i that player j will choose action a at h_j, then $t_i(\beta_i)$ must believe at h_i that player j is of some type $t_j(\beta_j)$ for which a is rational. In other words, every type $t_i(\beta_i)$ in $T_i(\mathcal{B})$ always believes in rationality at future information sets that are believed to be reached. However, since $t_i(\beta_i)$ believes at every information set that every opponent j is of some type $t_j(\beta_j)$ in $T_j(\mathcal{B})$, it follows that every $t_i(\beta_i)$ in $T_i(\mathcal{B})$ always believes in the event that all types believe in rationality at future information sets that are believed to be reached. By recursively applying this argument, one may conclude that every $t_i(\beta_i)$ in $T_i(\mathcal{B})$ respects common belief in the event that types always believe in rationality at future information sets that are believed to be reached. Quesada's sufficient condition can thus be formulated as follows in terms of our base model:

Quesada's condition BR: Type t_i respects common belief in the events that (1) types hold preferences as specified by $(\tilde{P}_j)_{j\in I}$, (2) types hold point-beliefs, and (3) types always believe in rationality at future information sets that are believed to be reached.

3.9 Samet's model

Samet presents a state-based semantic model which is an extension of the models by Aumann and Balkenborg & Winter [Sa$_2$96]. A Samet model is a tuple

$$\mathcal{M} = (\Omega, (B_i, f_i, v_i, \tau_i)_{i\in I}),$$

where Ω, B_i, f_i and v_i are as in the Aumann model, and τ_i is a so-called *hypothesis transformation* that assigns to every state ω and non-empty event $E \subseteq \Omega$ some new state ω'. My interpretation of τ_i is that if player i currently believes that the state is in $B_i(\omega)$, but later observes the event E, then he will believe that the state is in $B_i(\omega') \cap E$. Samet defines the hypothesis transformation in a different, but equivalent, way. In Samet's terminology, a hypothesis transformation assigns to every initial belief $B_i(\omega)$ and event E some new belief $B_i(\omega')$ for some $\omega' \in \Omega$. However, this definition is equivalent to the existence of a function τ_i as described in our model. The function τ_i must satisfy the following two conditions: (1) $B_i(\tau_i(\omega, E)) \cap E$ is nonempty for every ω and E, and (2) $\tau_i(\omega, E) = \omega$ whenever $B_i(\omega)$ has a nonempty intersection with E. These conditions indicate that $B_i(\tau_i(\omega, E)) \cap E$ may be interpreted as a well-defined conditional belief for player i at state ω when observing the event E.

As to the functions f_i, mapping states to strategy choices, it is assumed that for every terminal node z there is some state $\omega \in \Omega$ such that the profile $(f_i(\omega))_{i \in I}$ of stategies reaches z. This implies that for every information set h_i, the event

$$[h_i] = \{\omega \in \Omega \mid (f_i(\omega))_{i \in I} \text{ reaches } h_i\}$$

is nonempty, and hence can be used as a conditioning event for the hypothesis transformation τ_i. Samet assumes in his model a function ξ (instead of $(f_i)_{i \in I}$) mapping states to terminal nodes, and assumes that for every terminal node z there is some $\omega \in \Omega$ with $\xi(\omega) = z$. However, he shows that this function ξ induces, in some precise way, a profile $(f_i)_{i \in I}$ of strategy functions, as we use it. We work directly with the strategy functions here, in order to make the model as similar as possible to the Aumann model and the Balkenborg-Winter model.

In contrast to Aumann's model and Balkenborg and Winter's model, every state ω in Samet's model formally induces a conditional belief vector in our base model. Namely, take some state ω, a player i, and some information set $h_i \in H_i^*$. Then,

$$\hat{B}_i(\omega, h_i) := B_i(\tau_i(\omega, [h_i])) \cap [h_i]$$

respresents player i's conditional belief at h_i about the state. Since every state ω' induces for player j a strategy choice $f_j(\omega')$ and a conditional belief vector $(\hat{B}_j(\omega', h_j))_{h_j \in H_j^*}$, first-order conditional beliefs about the opponents' strategies, and higher-order conditional beliefs about the opponents' conditional beliefs can be derived at every state with the help of the hypothesis transformations τ_i. Hence, Samet's model can be expressed directly and formally in terms of our base model.

Samet's sufficient condition for backward induction is *common hypothesis of node rationality.* At state ω, player i said to be rational at $h_i \in H_i$ if (1) $\omega \in [h_i]$, and (2) there is no $s_i \in S_i$ such that for every $\omega' \in B_i(\omega) \cap [h_i]$ it holds that

$$v_i(\omega)(z(s_i, (f_j(\omega'))_{j \neq i} | h_i)) > v_i(\omega)(z(f_i(\omega), (f_j(\omega'))_{j \neq i} | h_i)),$$

where the definition of this expression is as in Aumann's model. Let $[\mathrm{rat}_i(h_i)]$ denote the set of states ω such that at ω, player i is rational at h_i. Common hypothesis of node rationality can now be defined by the following recursive procedure: For every player i and information set $h_i \in H_i^*$, let

$$\mathrm{CHNR}(h_i, h_i) = [\mathrm{rat}_i(h_i)].$$

Note that, by condition (1) above, $\mathrm{CHNR}(h_i, h_i)$ only contains states at which h_i is indeed reached. Now, let $k \geq 0$, and suppose that $\mathrm{CHNR}(h_i, h_j)$ has been defined for all information sets $h_i \in H_i^*, h_j \in H_j^*$ such that h_j comes after h_i, and there are at most k information sets between h_i and h_j. Suppose now that h_j comes after h_i, and that there are exactly $k + 1$ information sets between h_i and h_j. Let h be the unique information set that immediately follows h_i and precedes h_j. Define

$$\mathrm{CHNR}(h_i, h_j) = \{\omega \in \Omega \mid B_i(\tau_i(\omega, [h])) \cap [h] \subseteq \mathrm{CHNR}(h, h_j)\}.$$

Common hypothesis of node rationality is said to hold at state ω if $\omega \in \mathrm{CHNR}(h_0, h)$ for all information sets h. Hence, the player at h_0 believes that (1) every opponent j will choose rationally at those information sets h_j that immediately follow h_0, (2) every such opponent j will believe at every such h_j that every other player k will choose rationally at those h_k that immediately follow h_j, and so on.

Samet shows in Theorem 5.3 that for every profile $(v_i)_{i\in I}$ of utility functions, for every state ω at which common knowledge of $(v_i)_{i\in I}$ and common hypothesis of node rationality hold, the strategy profile $(f_i(\omega))_{i\in I}$ leads to the backward induction outcome with respect to $(v_i)_{i\in I}$. In particular, the player at h_0 chooses the backward induction action at h_0 with respect to $(v_i)_{i\in I}$. In Theorem 5.4, Samet shows that there always exists some state ω at which common knowledge of $(v_i)_{i\in I}$ and common hypothesis of node rationality hold.

For a given state ω and information set $h_i \in H_i^*$, say that common hypothesis of node rationality *at* h_i holds if $\omega \in \mathrm{CHNR}(h_i, h)$ for all information sets h that follow h_i. Then, Samet's Theorem 5.3 can be generalized as follows: For every $h_i \in H_i$ and every ω at which common knowledge of $(v_i)_{i\in I}$ and common hypothesis of node rationality at h_i hold, the strategy $f_i(\omega)$ chooses at h_i the backward induction action with respect to h_i.

In order to express this sufficient condition in terms of our base model, it is important to understand all implications of common hypothesis of node rationality. By definition, common hypothesis of node rationality at h_i implies that player i believes at h_i that (1) every opponent j will choose rationally at every information set h_j that immediately follows h_i, (2) every such opponent j will believe at every such h_j that every other player k will choose rationally at every h_k that immediately follows h_j, and so on. However, there are more implications.

Consider namely an information set $h_j \in H_j$ that immediately follows h_i and some information set $h_k \in H_k$ which immediately follows h_j such that $B_i(\tau_i(\omega, [h_j])) \subseteq [h_k]$. Hence, in terms of our base model, player i believes at h_i that h_k will be reached. Suppose that state ω is such that common hypothesis of node rationality at h_i holds at ω. By (1) above, it holds that (1') $B_i(\tau_i(\omega, [h_j])) \cap [h_j] \subseteq [\mathrm{rat}_j(h_j)]$. By (2) above, it holds for every $\omega' \in B_i(\tau_i(\omega, [h_j])) \cap [h_j]$ that (2') $B_j(\tau_j(\omega', [h_k])) \cap [h_k] \subseteq [\mathrm{rat}_k(h_k)]$. However, since $B_i(\tau_i(\omega, [h_j])) \subseteq [h_k]$, it follows that $\omega' \in [h_k]$ for every $\omega' \in B_i(\tau_i(\omega, [h_j]))$. Since $\omega' \in B_j(\omega')$, we have that $B_j(\omega')$ has a nonempty intersection with $[h_k]$, and hence (by the assumptions on τ_i) $\tau_j(\omega', [h_k]) = \omega'$ for every $\omega' \in B_i(\tau_i(\omega, [h_j]))$. We may therefore conclude that $B_j(\tau_j(\omega', [h_k])) = B_j(\omega')$ for every $\omega' \in B_i(\tau_i(\omega, [h_j])) \cap [h_j]$. By (2') it thus follows that $B_j(\omega') \cap [h_k] \subseteq [\mathrm{rat}_k(h_k)]$ for every $\omega' \in B_i(\tau_i(\omega, [h_j])) \cap [h_j]$. Since $\omega' \in B_j(\omega')$, and $\omega' \in [h_k]$ for every $\omega' \in B_i(\tau_i(\omega, [h_j]))$, it follows in particular that $\omega' \in [\mathrm{rat}_k(h_k)]$ for every $\omega' \in B_i(\tau_i(\omega, [h_j])) \cap [h_j]$, which means that player i believes at h_i that player k chooses rationally at h_k. Hence, we have shown that common hypothesis of node rationality at h_i implies that player i believes at h_i that player k chooses rationally at h_k whenever (1) there is only one information set between h_i and h_k, and (2) player i believes at h_i that h_k will be reached. By induction, one can now show that common hypothesis of node rationality at h_i implies that player i believes at h_i that player k chooses rationally at h_k whenever (1) h_k follows h_i and (2) player i believes at h_i that h_k can be reached.

By a similar argument, one can show that common hypothesis of node rationality at h_i implies that player i believes at h_i that common hypothesis of node rationality will hold at every future information set h_j which player i believes to be reached from h_i. Together with our previous insight, this means that common hypothesis of node rationality may be expressed, in terms of our base model, by forward belief in material rationality (see our Definition 2.7). Samet's sufficient condition for backward induction can thus be phrased as follows in terms of our base model:

Samet's condition BR: Type t_i (1) respects common belief in the event that types hold preferences as specified by $(\tilde{P}_j)_{j \in I}$, and (2) respects forward belief in material rationality.

3.10 Stalnaker's model

Stalnaker proposes a state-based semantic model for perfect information games in which every information set is controlled by a different player [St$_1$98]. The model we present here is not an exact copy of Stalnaker's model, but captures its essential properties. A Stalnaker model is a tuple

$$\mathcal{M} = (\Omega, (f_i, v_i, \lambda_i)_{i \in I})$$

where Ω, f_i and v_i are as in the Aumann model, and λ_i is a function that assigns to every state ω some lexicographic probability system (see Asheim's model) $\lambda_i(\omega)$ on Ω. That is, $\lambda_i(\omega)$ is a sequence $(\lambda_i^1(\omega), \ldots, \lambda_i^{K_i(\omega)}(\omega))$ where $\lambda_i^k(\omega)$ is a probability distribution on Ω. For every information set h let $[h] = \{\omega \in \Omega \mid (f_i(\omega))_{i \in I}$ reaches $h\}$. We assume that $[h]$ is non-empty for all h, and that $\lambda_i(\omega)$ has full support on Ω. By the latter, we mean that for every $\omega' \in \Omega$ there is some $k \in \{1, \ldots, K_i(\omega)\}$ such that $\lambda_i^k(\omega)$ assigns positive probability to ω'. As such, λ_i and $(f_j)_{j \neq i}$ induce, for every state ω, a probabilistic belief revision policy for player i in the following way. For every $h_i \in H_i^*$, let $k_i(\omega, h_i)$ be the first k such that $\lambda_i^k(\omega)$ assigns positive probability to $[h_i]$. Then, the probability distribution $\mu_i(\omega, h_i)$ on $[h_i]$ given by

$$\mu_i(\omega, h_i)(\omega') = \frac{\lambda_i^{k_i(\omega,h_i)}(\omega')}{\lambda_i^{k_i(\omega,h_i)}([h_i])}$$

for every $\omega' \in [h_i]$ represents player i's revised belief at ω upon observing that h_i has been reached. More generally, for every event $E \subseteq \Omega$, the probability distribution $\mu_i(\omega, E)$ on E given by

$$\mu_i(\omega, E)(\omega') = \frac{\lambda_i^{k_i(\omega,E)}(\omega')}{\lambda_i^{k_i(\omega,E)}(E)}$$

for every $\omega' \in E$ defines player i's revised belief upon receiving information E. Here, $k_i(\omega, E)$ is the first k such that $\lambda_i^k(\omega)$ assigns positive probability to E. The lexicographic probability system $\lambda_i(\omega)$ naturally induces, for every information set $h_i \in H_i^*$, the non-probabilistic conditional belief

$$\hat{B}_i(\omega, h_i) := \operatorname{supp} \mu_i(\omega, h_i),$$

and hence Stalnaker's model can be expressed directly in terms of our base model.

Stalnaker's sufficient condition for backward induction consists of *common initial belief in sequential rationality, and common belief in the event that players treat information about different players as epistemically independent.* Player i is called sequentially rational at ω if at every information set $h_i \in H_i^*$, the strategy $f_i(\omega)$ is optimal given the utility function $v_i(\omega)$ and the revised belief about the opponents' strategy choices induced by $\mu_i(\omega, h_i)$ and $(f_j)_{j \neq i}$. Let Ω^{srat} be the set of states at which all players are sequentially rational. Common initial belief in sequential rationality can be defined by the following recursive procedure:

$$\begin{aligned} \text{CIBSR}^1 &= \Omega^{\text{srat}}; \\ \text{CIBSR}^{k+1} &= \{\omega \in \Omega \mid \hat{B}_i(\omega, h_0) \subseteq \text{CIBSR}^k \text{ for all players } i\} \end{aligned}$$

for all $k \geq 1$. Common initial belief in sequential rationality is said to hold at ω if $\omega \in \text{CIBSR}^k$ for all k. We say that two states ω and ω' are indistinguishable for player i if $f_i(\omega) = f_i(\omega')$, $v_i(\omega) = v_i(\omega')$ and $\mu_i(\omega, h_i) = \mu_i(\omega', h_i)$ for all $h_i \in H_i^*$. An event E is said to be about player i if for every two states ω, ω' that are indistinguishable for player i, either both ω and ω'

are in E, or none is in E. We say that at ω player i treats information about different players as epistemically independent if for every two different opponents j and ℓ, for every event E_j about player j and every event E_ℓ about player ℓ, it holds that $\mu_i(\omega, E_j)(E_\ell) = \mu_i(\omega, \Omega \backslash E_j)(E_\ell)$ and $\mu_i(\omega, E_\ell)(E_j) = \mu_i(\omega, \Omega \backslash E_\ell)(E_j)$. In his theorem on page 43, Stalnaker shows that common initial belief in sequential rationality and common belief in the event that players treat information about different players as epistemically independent lead to backward induction.

In terms of our base model, common initial belief in sequential rationality corresponds to the condition that a type respects common initial belief in the event that types initially believe in rationality at all information sets. The epistemic independence condition cannot be translated that easily into our base model. The problem is that the base model only allows for beliefs conditional on *specific* events, namely events in which some information set is reached. On the other hand, in order to formalize the epistemic independence condition we need to condition beliefs on more general events. There is, however, an important consequence of the epistemic independence condition that can be expressed in terms of our base model, namely that the event of reaching information set h_i should not change player i's belief about the actions and beliefs of players that did not precede h_i. In order to see this, choose a player j that precedes h_i and a player ℓ that does not precede h_i. Note that the event of player j choosing the action leading to h_i is an event about player j, and that the event of player ℓ choosing a certain action and having a certain belief vector is an event about player ℓ. Hence, epistemic independence says that player i's belief about player ℓ's action and beliefs should not depend on whether player j has moved the game towards h_i or not. Moreover, it is exactly this consequence of epistemic independence that drives Stalnaker's backward induction result. In particular, if player i initially believes that player ℓ chooses rationally at his information set h_ℓ (which does not precede h_i), then player i should continue to believe so if he observes that h_i has been reached. If we drop the assumption that every player only controls one information set, the condition amounts to saying that a player should never revise his belief about the actions and beliefs at future and parallel information sets.

In terms of our base model, Stalnaker's sufficient condition for backward induction can thus be stated as follows:

Stalnaker's condition BR: Type t_i respects common belief in the events that (1) types hold preferences as specified by $(\tilde{P}_j)_{j \in I}$, and (2) types do not change their belief about the opponents' choices and beliefs at future and parallel information sets, and type t_i respects common initial belief in the event that (3) types initially believe in rationality at all information sets.

Halpern provides an explicit comparison between the models of Aumann and Stalnaker [Ha$_0$01].

	Ash	A&P	Aum	B&W	Cla	Fei1	Fei2	Per	Que	Sam	Sta
Common belief in event that types. . .											
. . . initially believe in rat. at all information sets			●								
. . . always believe in rat. at future information sets that are believed to be reached									●		
. . . always believe in rat. at all future information sets	●										
. . . always believe in rat. at all future and parallel information sets		●				●					
. . . always believe in rat. at all information sets								●			
Common initial belief in event that types. . .											
. . . initially believe in rat. at all inf. sets											●
Forward belief in. . .											
. . . substantive rationality				●	●		●				
. . . material rationality										●	
Common belief in event that types. . .											
. . . never revise belief about opponents' preference relations	●	●	●	●	●	●	●		●	●	●
. . . do not revise belief about opponents' choices and beliefs at future and parallel information sets			●	●							●
. . . minimally revise belief about opponents' preferences and beliefs								●			
. . . hold point-beliefs								●	●		

Table 1. Overview of sufficient conditions for backward induction

3.11 Summary

The discussion of the various models and sufficient conditions for backward induction can be summarized by Table 1.

The table shows that several sufficient conditions for backward induction, although formulated in completely different epistemic models, become equivalent once they have been expressed in terms of our base model. Note also that there is no model assuming common belief in the events that (1) types always believe in rationality at *all* information sets, and (2) types never revise their beliefs about the opponents' preferences over terminal nodes. This is no surprise, since the papers [Re$_3$92, Re$_3$93] have illustrated that these two events are in general incompatible. Perea's model maintains condition (1) and weakens condition (2), while the other models maintain condition (2) and weaken condition (1). Finally observe that all models assume (at least) initial common belief in the event that types initially believe in rationality at all information sets, plus some extra conditions on the players' belief revision procedures. If one would only assume the former, this would lead to the concept of *common certainty of rationality at the beginning of the game*, as defined in [BP97]. This concept is considerably weaker than backward induction, as it may not even lead to the backward induction outcome. Hence, additional conditions on the players' belief revision policies are needed in each model to arrive at backward induction.

References

[As02] G.B. Asheim. On the epistemic foundation for backward induction. *Mathematical Social Sciences* 44(2):121–144, 2002.

[AsPe$_2$05] G.B. Asheim & A. Perea. Sequential and quasi-perfect rationalizability in extensive games. *Games and Economic Behavior* 53(1):15–42, 2005.

[Au95] R.J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior* 8(1):6–19, 1995.

[Au98] R.J. Aumann. On the centipede game. *Games and Economic Behavior* 23(1):97–105, 1998.

[Ba$_3$Wi$_2$97] D. Balkenborg & E. Winter. A necessary and sufficient epistemic condition for playing backward induction. *Journal of Mathematical Economics* 27(3):325–345, 1997.

[Ba$_7$97] P. Battigalli. On rationalizability in extensive games. *Journal of Economic Theory* 74(1):40–61, 1997.

[Ba$_7$Si$_1$02] P. Battigalli & M. Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory* 106(2):356–391, 2002.

[BP97] E. Ben-Porath. Rationality, Nash equilibrium and backwards induction in perfect-information games. *Review of Economic Studies* 64(1):23–46, 1997.

[Bi$_1$87] K. Binmore. Modeling rational players, part I. *Economics and Philosophy* 3:179–214, 1987.

[Bo$_0$92] G. Bonanno. Rational beliefs in extensive games. *Theory and Decision* 33(2):153–176, 1992.

[Br$_1$Fr$_2$Ke$_1$04] A. Brandenburger, A. Friedenberg & H.J. Keisler. Admissibility in games, 2004. Forthcoming.

[Br$_3$Ra$_0$99] J. Broome & W. Rabinowicz. Backwards induction in the centipede game. *Analysis* 59(264):237–242, 1999.

[Ca$_2$00] J.W. Carroll. The backward induction argument. *Theory and Decision* 48(1):61–84, 2000.

[Cl$_0$03] T. Clausing. Doxastic conditions for backward induction. *Theory and Decision* 54(4):315–336, 2003.

[Cl$_0$04] T. Clausing. Belief revision in games of perfect information. *Economics and Philosophy* 20:89–115, 2004.

[Fe$_1$05] Y. Feinberg. Subjective reasoning—dynamic games. *Games and Economic Behavior* 52(1):54–93, 2005.

[Ha$_0$01] J.Y. Halpern. Substantive rationality and backward induction. *Games and Economic Behavior* 37(2):425–435, 2001.

[Ho$_0$Lo$_3$13] E.W. Hobson & A.E.H. Love, eds. *Proceedings of the Fifth International Congress of Mathematicians, Vol. 2.* Cambridge University Press, 1913.

[Pe$_0$84] D.G. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52(4):1029–1050, 1984.

[Pe$_2$05] A. Perea. Minimal belief revision leads to backward induction, 2005. Unpublished.

[Pr00] G. Priest. The logic of backward inductions. *Economics and Philosophy* 16:267–285, 2000.

[Qu02] A. Quesada. Belief system foundations of backward induction. *Theory and Decision* 53(4):393–403, 2002.

[Qu03] A. Quesada. From common knowledge of rationality to backward induction. *International Game Theory Review* 5(2):127–137, 2003.

[$\mathrm{Ra}_0$98] W. Rabinowicz. Grappling with the centipede. *Economics and Philosophy* 14:95–126, 1998.

[$\mathrm{Re}_3$92] P.J. Reny. Rationality in extensive-form games. *Journal of Economic Perspectives* 6(4):103–118, 1992.

[$\mathrm{Re}_3$93] P.J. Reny. Common belief and the theory of games with perfect information. *Journal of Economic Theory* 59(2):257–274, 1993.

[$\mathrm{Ro}_3$81] R.W. Rosenthal. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory* 25(1):92–100, 1981.

[$\mathrm{Ru}_1$91] A. Rubinstein. Comments on the interpretation of game theory. *Econometrica* 59(4):909–924, 1991.

[$\mathrm{Sa}_2$96] D. Samet. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior* 17(2):230–251, 1996.

[$\mathrm{St}_1$96] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy* 12:133–163, 1996.

[$\mathrm{St}_1$98] R. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences* 36(1):31–56, 1998.

[Ze13] E. Zermelo. Über eine Anwendung der Mengenlehre auf die Theorie des Schachpiel. In [$\mathrm{Ho}_0\mathrm{Lo}_3$13, pp. 501–504].