# Behavior and deliberation in perfect-information games: Nash equilibrium and backward induction

**Giacomo Bonanno**

*Department of Economics, University of California, Davis, USA*
gfbonanno@ucdavis.edu

**Abstract**

Doxastic characterizations of the set of Nash equilibrium outcomes and of the set of backward-induction outcomes are provided for general perfect-information games (where there may be multiple backward-induction solutions). We use models that are behavioral, rather than strategy-based, where a state only specifies the actual play of the game and not the hypothetical choices of the players at nodes that are not reached by the actual play. The analysis is completely free of counterfactuals and no belief revision theory is required, since only the beliefs at reached histories are specified.

## 1 Introduction

We provide a doxastic (that is, belief-based, rather than knowledge-based) characterization of the set of Nash equilibrium outcomes and of the set of backward-induction outcomes for *general* finite perfect-information games: previous characterizations were provided only for games with no relevant ties or generic games.[1] We make use of models that are behavioral, rather than strategy-based, in the sense that a state only specifies the actual play of the game and not the hypothetical choices of the players at counterfactual nodes, that is, nodes that are not reached by the actual play. Our analysis is completely free of counterfactuals (both objective and subjective), as explained below.

We use the history-based definition of extensive-form game (see, for example Osborne and Rubinstein (1994)); details are provided in Section 2. If $h$ and $h'$ are two histories we denote by $h \prec h'$ the fact that $h$ is a *proper* prefix of $h'$,[2] while $h \preceq h'$ means that $h$ is a prefix of $h'$, that is, either $h \prec h'$ or $h = h'$. The set of decision histories is denoted by $D$ and the set of terminal histories by $Z$.

The following are the main features of our approach.

1. The models that we use are *behavioral* models where a state specifies the *actual* sequence of moves. Strategies play no role.[3] We denote the set of states by $\Omega$ and use a function $\zeta : \Omega \to Z$ to specify, for every state $\omega \in \Omega$, the terminal history $\zeta(\omega)$ (that is, the play of the game) associated with $\omega$.

2. For every state $\omega$ we specify only the *actual* beliefs of the relevant player at every decision history that is *actually reached* at $\omega$ (that is, for decision histories $h$ such that $h \prec \zeta(\omega)$). No objective or subjective counterfactuals are postulated. Furthermore, no belief revision theory is needed, since the models that we use do not specify "initial" beliefs nor do they rely on any restriction about how the beliefs of a player evolve along a given play (should a player move more than once along that play).

3. In line with the philosophy literature that emphasizes that "the deliberating agent cannot, before choice, predict how he will choose" (Levi 1997, p.65),[4] at every reached decision history we endow the active player with a belief about what will happen if she takes action $a$, *for every available action $a$*. Thus the beliefs that we model are "pre-choice" or "deliberation-stage" beliefs. This is a departure from the standard approach in the game-theoretic literature where it is assumed that, at every state, if a player takes a particular action then she knows that she takes that action.

[1] See, for example, Aumann (1995), Balkenborg and Winter (1997), Ben-Porath (1997), Bonanno (2013), Clausing (2003), Halpern (2001), Perea (2012; 2014), Quesada (2003), Samet (1996; 2013), Stalnaker (1998). Surveys of the literature on the epistemic foundations of backward induction are provided in Brandenburger (2007), Perea (2007a) and (Perea 2012, p.463).

[2] If one identifies histories with nodes in the tree, then $h \prec h'$ means that node $h$ is a predecessor of node $h'$.

[3] Behavioral models were first introduced in Samet (1996).

[4] See also Gilboa (1999), Ginet (1962), Goldman (1970), Ledwig (2005), Spohn (1977; 1999).

4. We use a very weak notion of rationality, which has been referred to in the literature as "material rationality".[5] First of all, for every state $\omega$, rationality is only evaluated at decision histories that are actually reached at $\omega$ (and only for the active players at those decision histories). Secondly, if $h$ is a decision history that is reached at state $\omega$, the player who is active at $h$ is rational if the action that she actually takes at $h$ (at state $\omega$) is optimal given her beliefs, in the sense that it is not the case that - according to her beliefs - there is another action of hers that *guarantees* higher utility.

The first result (Proposition 1) provides the following characterization of the set of Nash equilibrium outcomes:[6]

- Given a perfect-information game and an arbitrary model of it, if $\omega$ is a state where, at every reached history, (1) no player has false beliefs, (2) every player is rational and (3) no player has uncertainty about what will happen after any of her choices, then $\zeta(\omega)$ - the terminal history associated with $\omega$ - is a Nash equilibrium play (that is, there is a Nash equilibrium whose associated play is $\zeta(\omega)$).
- If $z$ is a terminal history generated by a Nash equilibrium, then there is a model of the game and a state $\omega$ in that model such that ($a$) $\zeta(\omega) = z$ and ($b$) at $\omega$ and at every reached history (1) no player has false beliefs, (2) every player is rational and (3) no player has uncertainty about what will happen after any of her choices.

The above conditions under (1)-(3) are expressed as events, denoted by **T**, **R** and **C**, respectively (T for 'Truth', R for 'Rationality' and C for 'Certainty'). Thus the set of Nash equilibrium outcomes is characterized by the event $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$.

The second result (Proposition 2) provides a characterization of the set of backward-induction outcomes in terms of a strengthening of the above conditions, obtained by intersecting the event $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$ with one more event, $\mathbf{I}_{TRC}$ (where I stands for 'Iterated'), which expresses the following conditions: the root player believes that if she takes action $a$ and $a$ is a decision history then (1) the active player at $a$ has correct beliefs, is rational and has no uncertainty, (2) the active player at $a$ believes that if he takes action $b$ and $ab$ is a decision history then the active player at $ab$ has correct beliefs, is rational and has no uncertainty,

[5] See, for example, Aumann (1995; 1998), Battigalli et al. (2013), Samet (1996).

[6] For simplicity, the characterization is provided for games where no player moves more than once along any play, but we explain how to extend the result to general games. The words 'outcome', 'play' and 'terminal history' will be used interchangeably.

(3) the active player at $a$ believes that the active player at $ab$ believes that if she takes action $c$ and $abc$ is a decision history then the active player at $abc$ has correct beliefs, is rational and has no uncertainty, and so forth. This condition can be expressed using belief operators as explained in Section 5.1. Proposition 2 provides the following characterization of the set of backward-induction outcomes:[7]

- Given a perfect-information game and an arbitrary model of it, if $\omega \in (\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC}$ then $\zeta(\omega)$ - the terminal history associated with $\omega$ - is a backward-induction outcome.

- If the terminal history $z$ is a backward-induction outcome, then there is a model of the game and a state $\omega$ in that model such that $(a)$ $\zeta(\omega) = z$ and $(b)$ $\omega \in (\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC}$.

The paper is organized as follows. The next section introduces the class of behavior-deliberation models, Section 3 defines the notion of rationality, Section 4 provides the characterization of Nash equilibrium outcomes, Section 5 characterizes the set of backward-induction outcomes, Section 6 examines related literature and Section 7 discusses a number of conceptual issues that are raised by the approach put forward in this paper. The proofs are given in the Appendix.

## 2 Behavioral models of perfect-information games

We use the history-based definition of extensive-form game, which is as follows. If $A$ is a set, we denote by $A^*$ the set of finite sequences in $A$. If $h = \langle a_1, ..., a_k \rangle \in A^*$ and $1 \leq j < k$, the sequence $\left\langle a_1, ..., a_j \right\rangle$ is called a *proper prefix* of $h$. We denote the fact that $h'$ is a proper prefix of $h$ by $h' \prec h$, while $h' \preceq h$ means that either $h' \prec h$ or $h' = h$. If $h = \langle a_1, ..., a_k \rangle \in A^*$ and $a \in A$, we denote the sequence $\langle a_1, ..., a_k, a \rangle \in A^*$ by $ha$.

**Definition 2.1.** A *finite extensive form with perfect information* (without chance moves) is a tuple $\langle A, H, N, \iota \rangle$ whose elements are:

- A finite set of *actions* $A$ and a finite set of *histories* $H \subseteq A^*$ which is closed under prefixes (that is, if $h \in H$ and $h' \in A^*$ is such that $h' \prec h$, then

[7] This characterization is *not* restricted to games where no player moves more than once along any play.

$h' \in H$). The null history $\langle\rangle$, denoted by $\emptyset$, is an element of $H$ and is a prefix of every history. A history $h \in H$ such that, for every $a \in A$, $ha \notin H$, is called a *terminal history*. The set of terminal histories is denoted by $Z$. $D = H \setminus Z$ denotes the set of non-terminal or *decision* histories. For every history $h \in D$, we denote by $A(h)$ the set of actions available at $h$, that is, $A(h) = \{a \in A : ha \in H\}$.

- A finite set $N$ of *players* and a function $\iota : D \to N$ that assigns a player to each decision history. Thus $\iota(h)$ is the player who moves at history $h$; we refer to that player as *the active player at history h*. For every $i \in N$, $D_i$ denotes the set of decision histories of player $i$, that is, $D_i = \{h \in D : \iota(h) = i\}$.

Given an extensive form, one obtains an *extensive game* by adding, for every player $i \in N$, a *utility* (or *payoff*) *function* $u_i : Z \to \mathbb{R}$ (where $\mathbb{R}$ denotes the set of real numbers; recall that $Z$ is the set of terminal histories).

From now on, histories will be denoted more succinctly by listing the corresponding actions, without angled brackets, without commas and omitting the null history: thus instead of writing $\langle\emptyset, a_1, a_2, a_3, a_4\rangle$ we will simply write $a_1a_2a_3a_4$.

Before introducing the definition of a model of a game, we recall the following facts about belief relations and operators. If $\Omega$ is a set (whose elements are called "states") and $\mathcal{B} \subseteq \Omega \times \Omega$ is a binary relation on $\Omega$ (representing the beliefs of an individual), for every $\omega \in \Omega$ we denote by $\mathcal{B}(\omega)$ the set of states that are reachable from $\omega$ using $\mathcal{B}$, that is, $\mathcal{B}(\omega) = \{\omega' \in \Omega : \omega\mathcal{B}\omega'\}$.[8] $\mathcal{B}$ is *serial* if $\mathcal{B}(\omega) \neq \varnothing$, for every $\omega \in \Omega$; it is *transitive* if $\omega' \in \mathcal{B}(\omega)$ implies $\mathcal{B}(\omega') \subseteq \mathcal{B}(\omega)$ and it is *euclidean* if $\omega' \in \mathcal{B}(\omega)$ implies $\mathcal{B}(\omega) \subseteq \mathcal{B}(\omega')$. Subsets of $\Omega$ are called "events". Given an event $E \subseteq \Omega$, we say that at $\omega \in \Omega$ the individual believes $E$ if and only if $\mathcal{B}(\omega) \subseteq E$. Thus one can define a *belief operator* $\mathbb{B} : 2^\Omega \to 2^\Omega$ as follows: $\mathbb{B}E = \{\omega \in \Omega : \mathcal{B}(\omega) \subseteq E\}$; hence $\mathbb{B}E$ is the event that the individual believes $E$. It is well known that seriality of $\mathcal{B}$ corresponds to consistency of beliefs (if the individual believes $E$ then it is not the case that she believes not $E : \mathbb{B}E \subseteq \neg\mathbb{B}\neg E$, where, for every event $F$, $\neg F$ denotes the complement of $F$ in $\Omega$), transitivity corresponds to positive introspection (if the individual believes $E$ then she believes that she believes $E : \mathbb{B}E \subseteq \mathbb{B}\mathbb{B}E$) and euclideanness corresponds to negative introspection (if the individual does not believe $E$ then she believes that she does not believe $E : \neg\mathbb{B}E \subseteq \mathbb{B}\neg\mathbb{B}E$).[9]

[8] As is customary, we take $\omega\mathcal{B}(\omega')$ and $(\omega, \omega') \in \mathcal{B}$ as interchangeable.

[9] For more details see Battigalli and Bonanno (1999).

To define a model of a game, we begin with a set $\Omega$, whose elements are called *states* and whose subsets are called *events*. We interpret each state in terms of a particular complete play of the game, by means of a function $\zeta : \Omega \to Z$ that associates, with every state $\omega$, a terminal history $\zeta(\omega) \in Z$. Next we add, for every decision history $h \in D$, a binary relation $\mathcal{B}_h$ on $\Omega$ representing the beliefs of $\iota(h)$, the active player at $h$;[10] however, we do so only at histories that are actually reached at a given state, in the sense that $\mathcal{B}_h(\omega) \neq \emptyset$ if and only if $h \prec \zeta(\omega)$.

**Definition 2.2.** Given a perfect-information game, a *model of it* is a tuple $\mathcal{M} = \langle \Omega, \zeta, \{\mathcal{B}_h\}_{h \in D} \rangle$ where

- $\Omega$ is a set of states.
- $\zeta : \Omega \to Z$.
- For every $h \in D$, $\mathcal{B}_h \subseteq \Omega \times \Omega$ is a belief relation that satisfies the following properties:
    1. $\mathcal{B}_h(\omega) \neq \emptyset$ if and only if $h \prec \zeta(\omega)$ [beliefs are specified only at reached decision histories and are consistent].
    2. If $\omega' \in \mathcal{B}_h(\omega)$ then $\mathcal{B}_h(\omega') = \mathcal{B}_h(\omega)$ [beliefs satisfy positive and negative introspection].
    3. If $\omega' \in \mathcal{B}_h(\omega)$ then $h \prec \zeta(\omega')$ [the active player at history $h$ knows that $h$ has been reached].
    4. If $\mathcal{B}_h(\omega) \neq \emptyset$ then, for every action $a \in A(h)$, there is an $\omega' \in \mathcal{B}_h(\omega)$ such that $ha \preceq \zeta(\omega')$ .

The last condition states that, for *every* action $a$ available at $h$, there is a state $\omega'$ that the active player at $h$ considers possible ($\omega' \in \mathcal{B}_h(\omega)$) where she takes action $a$ (that is, history $ha$ is a prefix of $\zeta(\omega')$). This means that, for every available action, the active player at $h$ has a belief about what will, or might, happen if she chooses that action. Note that this way of modeling beliefs is a departure from the standard approach in the literature, where it is assumed that if, at a state, a player takes a particular action then she knows that she takes that action. The standard approach thus requires the use of either objective or subjective counterfactuals in order to represent a player's beliefs about the consequences of taking alternative actions.[11] In our approach

[10] Thus it would be more precise to write $\mathcal{B}_{\iota(h)}$ instead of $\mathcal{B}_h$, but we have chosen the lighter notation since there is no ambiguity, because at every decision history there is a unique player who is active there.

[11] For a critical analysis of the use of counterfactuals in dynamic games see Bonanno (2015).

a player's beliefs refer to the *deliberation* or *pre-choice stage*, where the player considers the consequences of taking any available action, without pre-judging her subsequent decision.[12]

Since the state encodes the player's actual choice, that choice can be judged to be rational or irrational by relating it to the player's pre-choice beliefs. Thus it is possible for a player to have the same beliefs at two different states, say $\alpha$ and $\beta$, and be labeled as rational at state $\alpha$ and irrational at state $\beta$, because the action she ends up taking at state $\alpha$ is optimal given those beliefs, while the action she ends up taking at state $\beta$ is not optimal given those same beliefs. The formal definition of rationality is given in Section 3.

Consider the game shown in Figure 1, together with a model of it.[13] We represent a belief relation $\mathcal{B}$ as follows: for any two states $\omega$ and $\omega'$, $\omega' \in \mathcal{B}(\omega)$ if and only if either $\omega$ and $\omega'$ are enclosed in the same rounded rectangle or there is an arrow from $\omega$ to the rounded rectangle containing $\omega'$.[14] The relations shown in the model of Figure 1 are those of the active players: the relation at the null history $\emptyset$, $\mathcal{B}_\emptyset$, is that of Player 1, the relation at history $a_1$, $\mathcal{B}_{a_1}$, is that of Player 2, etc.[15]

Consider a state, say $\delta$. Then $\delta$ describes the following beliefs: at the null history $\emptyset$ (the root of the tree) the active player (Player 1) believes that if she takes action $a_1$ then Player 2 will either follow with action $d_2$ (state $\beta$) or with action $a_2$ followed by action $d_3$ of Player 3 (state $\gamma$) and if she takes action $d_1$ then the play will end (state $\alpha$); at history $a_1$ the active player (Player 2) knows that Player 1 played $a_1$ and believes that if he takes action $a_2$ then Player 3 will follow with $d_3$ (state $\gamma$) and if he takes action $d_2$ the play will end (state $\beta$); at history $a_1a_2$ the active player (Player 3) knows that Players 1 and 2 played $a_1$ and $a_2$, respectively, and believes that if she takes action $a_3$ then Player 1 will follow with $d_4$ (state $\delta$) and if she takes action $d_3$ then the play will end (state $\gamma$), etc. At state $\delta$, Player 1 ends up playing $a_1$ (at the root of the tree), Player 2 ends up playing $a_2$, Player 3 ends up playing $a_3$ and Player 1, at her last move, ends up playing $d_4$ .

[12] This issue is further discussed in Section 7.3.

[13] The root of the tree corresponds to the null history $\emptyset$, Player 2's decision node corresponds to history $a_1$, Player 3's decision node to history $a_1a_2$ and Player 1's last decision node to history $a_1a_2a_3$.

[14] In other words, for any two states $\omega$ and $\omega'$ that are enclosed in a rounded rectangle, $\{(\omega,\omega),(\omega,\omega'),(\omega',\omega),(\omega',\omega')\} \subseteq \mathcal{B}$ (that is, the relation is total on the set of states contained in the rectangle) and if there is an arrow from a state $\omega$ to a rounded rectangle then, for every $\omega'$ in the rectangle, $(\omega,\omega') \in \mathcal{B}$.

[15] Thus $\mathcal{B}_\emptyset(\omega) = \{\alpha,\beta,\gamma\}$ for every $\omega \in \Omega = \{\alpha,\beta,\gamma,\delta,\epsilon\}$, $\mathcal{B}_{a_1}(\omega) = \{\beta,\gamma\}$ for every $\omega \in \{\beta,\gamma,\delta,\epsilon\}$, $\mathcal{B}_{a_1a_2}(\omega) = \{\gamma,\delta\}$ for every $\omega \in \{\gamma,\delta,\epsilon\}$ and $\mathcal{B}_{a_1a_2a_3}(\omega) = \{\delta,\epsilon\}$ for every $\omega \in \{\delta,\epsilon\}$.

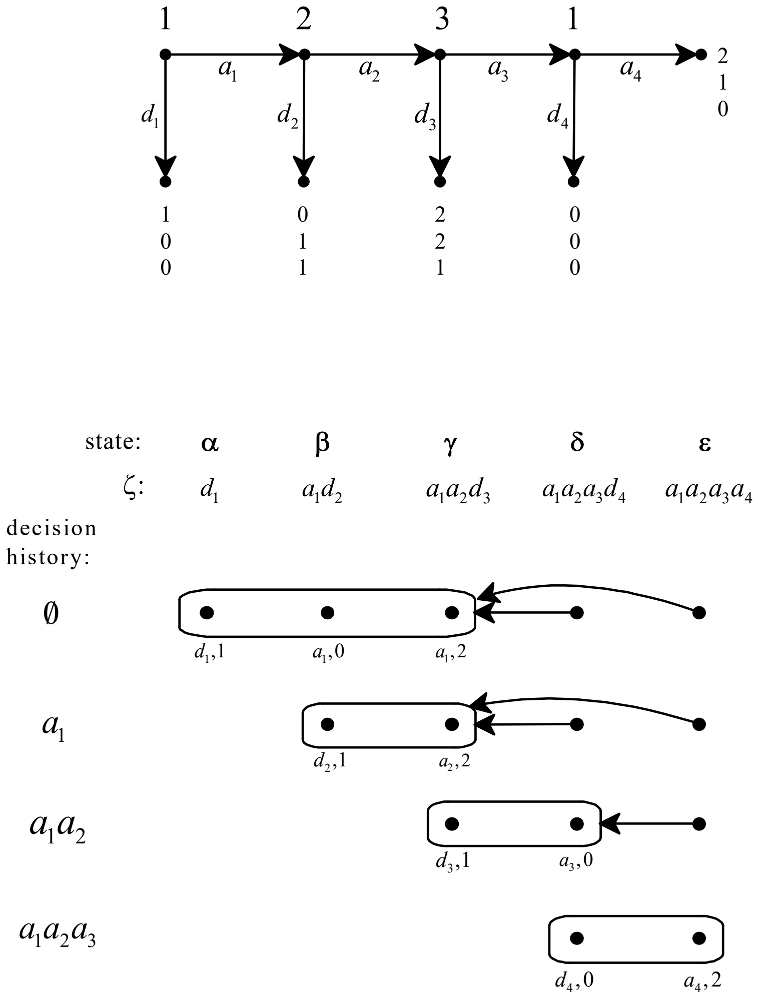

Figure 1: A perfect-information game and a model of it

In the model of Figure 1, under each state that is considered possible (by the relevant player) we have recorded the action actually taken by the player and that player's payoff (at the terminal history associated with that state). This will be useful when assessing the rationality of a player at a state. The issue of rationality is addressed in the next section.

**Remark 1.** *It is worth stressing that the notion of model that we are using allows for erroneous beliefs. For example, in the model of Figure 1, at state $\delta$ and history $a_1$ Player 2 has incorrect beliefs about the subsequent move of Player 3 if she herself plays $a_2$: she believes that Player 3 will follow with $d_3$ while, in fact, Player 3 plays $a_3$.*

# 3 Rationality

We use a very weak notion of rationality, which has been referred to in the literature as "material rationality" (see, for example, Aumann (1995; 1998), Battigalli et al. (2013), Samet (1996)). We say that, at state $\omega$ and at a decision history $h$ that is reached at $\omega$, the active player is rational if her actual action (at $h$ and $\omega$) is "optimal" given her beliefs, in the sense that it is not the case that - according to her beliefs - there is another action of hers that *guarantees* higher utility.[16]

**Definition 3.1.** Let $\omega$ be a state, $h$ a decision history that is reached at $\omega$ (that is, $h \prec \zeta(\omega)$) and $a, b \in A(h)$ two actions available at $h$. We say that, at $\omega$ and $h$, the active player $\iota(h)$ *believes that $b$ is better than $a$* if, $\forall \omega_1, \omega_2 \in \mathcal{B}_h(\omega)$, if $ha \preceq \zeta(\omega_1)$ (that is, $a$ is the action taken at history $h$ at state $\omega_1$) and $hb \preceq \zeta(\omega_2)$ (that is, $b$ is the action taken at history $h$ at state $\omega_2$) then $u_{\iota(h)}(\zeta(\omega_1)) < u_{\iota(h)}(\zeta(\omega_2))$ (recall that, for every player $i$, $u_i : Z \to \mathbb{R}$ is player $i$'s utility function on the set of terminal histories). In other words, the active player at $h$ believes that $b$ is better than $a$ if, restricting attention to the states that she considers possible, the *maximum* utility that she obtains if she plays $a$ is less than the *minimum* utility that she obtains if she plays $b$.

For example, in the model shown in Figure 1, at state $\delta$ and at the null history $\emptyset$ it is not the case that the active player (Player 1) believes that action $a_1$

[16] Note that rationality in the traditional sense of expected utility maximization implies rationality in our sense; thus anything that is implied by our weak notion will also be implied by the stronger notion of expected utility maximization. On the other hand, our notion has the advantage that it does not rely on the assumption of von Neumann-Morgenstern preferences: the utility functions can be just *ordinal* utility functions.

is better than action $d_1$ (since $\beta \in \mathcal{B}_\emptyset(\delta), a_1 \prec \zeta(\beta) = a_1d_2, \alpha \in \mathcal{B}_\emptyset(\delta), d_1 \preceq \zeta(\alpha) = d_1$ and $u_1(a_1d_2) = 0 < u_1(d_1) = 1$) and it is also not the case that Player 1 believes that action $d_1$ is better than action $a_1$ (since $\gamma \in \mathcal{B}_\emptyset(\delta), a_1 \prec \zeta(\gamma) = a_1a_2d_3$, $\alpha \in \mathcal{B}_\emptyset(\delta), d_1 \preceq \zeta(\alpha) = d_1$ and $u_1(d_1) = 1 < u_1(a_1d_2d_3) = 2$); in other words, at state $\delta$ Player 1 believes that if she plays $d_1$ her utility will be 1 and if she plays $a_1$ her utility might be 0 or might be 2. On the other hand, at state $\delta$ and at decision history $a_1$ the active player (Player 2) believes that action $a_2$ is better than action $d_2$.

Using Definition 3.1 we can define the event that the active player is rational at decision history $h$; we denote this event by $\mathbf{R}_h$.

**Definition 3.2.** Let $h$ be a decision history and $\omega$ a state. Then $\omega \in \mathbf{R}_h$ if and only if (1) $h \prec \zeta(\omega)$ and (2) if $ha \preceq \zeta(\omega)$ (that is, $a \in A(h)$ is the action played at $h$ at state $\omega$) then, for every $b \in A(h)$, it is not the case that, at state $\omega$ and history $h$, player $\iota(h)$ believes that $b$ is better than $a$.

For example, in the model of Figure 1 we have that $\mathbf{R}_\emptyset = \Omega$, $\mathbf{R}_{a_1} = \{\gamma, \delta, \epsilon\}$ (Player 2 believes that $a_2$ is better than $d_2$ and thus he is rational at, and only at, those states where he plays $a_2$), $\mathbf{R}_{a_1a_2} = \{\gamma\}$ and $\mathbf{R}_{a_1a_2a_3} = \{\epsilon\}$.

**Definition 3.3.** Let $\mathbf{R}$ be the event that at every *reached* decision history the active player is rational: $\omega \in \mathbf{R}$ if and only if $\omega \in \mathbf{R}_h$ for all $h \in D$ such that $h \prec \zeta(\omega)$.

For example, in the model of Figure 1 we have that $\mathbf{R} = \{\alpha, \gamma\}$.

# 4 Nash equilibrium play

Suppose that we have a model of a perfect-information game where there is a state $\omega$ such that, at every reached decision history, the active player is rational (that is, $\omega \in \mathbf{R}$: see Definition 3.3); what can we say about $\zeta(\omega)$, the actual play at that state? In general, we cannot conclude that $\zeta(\omega)$ is a Nash equilibrium outcome. To see this, consider the game and model shown in Figure 2, where $\mathbf{R} = \{\alpha, \beta\}$.[17] Thus $\alpha \in \mathbf{R}$ and yet $\zeta(\alpha) = a_1b_2$ which is an outcome that cannot be generated by a Nash equilibrium (that is, there is no Nash equilibrium whose associated outcome is $a_1b_2$, since - when Player 2's strategy is to play $b_2$ - Player 1's payoff would increase from 0 to 1 if she switched her choice from $a_1$ to $a_2$). In this model, at state $\alpha$ Player 1 erroneously believes that if she plays $a_1$ then

[17] In fact, $\mathbf{R}_\emptyset = \mathbf{R}_{a_1} = \{\alpha, \beta\}$.

Player 2 will follow with $b_1$ ($\beta \in \mathcal{B}_{\emptyset}(\alpha)$ and $\zeta(\beta) = a_1 b_1$) while, as a matter of fact, Player 2 plays $b_2$.

In order to obtain a characterization of Nash equilibrium outcomes we need to rule out erroneous beliefs.

It is well-known that, in general, correctness of beliefs corresponds to the property of reflexivity of the belief relations, which in our case would be expressed as follows: $\forall \omega \in \Omega, \forall h \in D$, if $h \prec \zeta(\omega)$ then $\omega \in \mathcal{B}_h(\omega)$. However, there are two reasons why one should *not* assume the belief relations to be reflexive: the first is a general conceptual reason and the second is a reason specific to our class of models.

The conceptual reason is that when the belief relations are assumed to be reflexive, beliefs become *necessarily* correct (and one can speak of *knowledge* rather than belief). As Stalnaker (Stalnaker (1996)) points out, it is methodologically preferable to carry out the analysis in terms of (possibly erroneous) beliefs and then - if desired - add further conditions that are sufficient to rule out incorrect beliefs *at a particular state*. The reason why one should not start with the assumption of necessarily correct beliefs (that is, reflexivity of the belief relations) is that this assumption has strong intersubjective implications:

> "The assumption that Alice believes (with probability one) that Bert believes (with probability one) that the cat ate the canary tells us nothing about what Alice believes about the cat and the canary themselves. But if we assume instead that Alice knows that Bert knows that the cat ate the canary, it follows, not only that the cat in fact ate the canary, but that Alice knows it, and therefore believes it as well." [Stalnaker (1996), p. 153.]

Thus we want a weaker notion of correctness of beliefs which, for example, allows for the possibility that a player has correct beliefs about the beliefs of another player without subscribing to those beliefs, that is, without at the same time believing that the beliefs of that player are in fact correct.

The second reason why one should not assume reflexivity of the belief relations is specific to our structures: reflexivity would imply complete uncertainty in the mind of each active player as to what will happen if she chooses alternative actions. This is due to the fact that, by Point 4 of Definition 2.2, for every action available at a decision history $h$, there should be a state where player $\iota(h)$ (the active player at $h$) takes that action; if player $\iota(h)$ happens to believe that after taking, say, action $a$ the following player will play, say, action $b$ rather than action $c$, then at a state where, as a matter of fact, action $c$ is played player $\iota(h)$

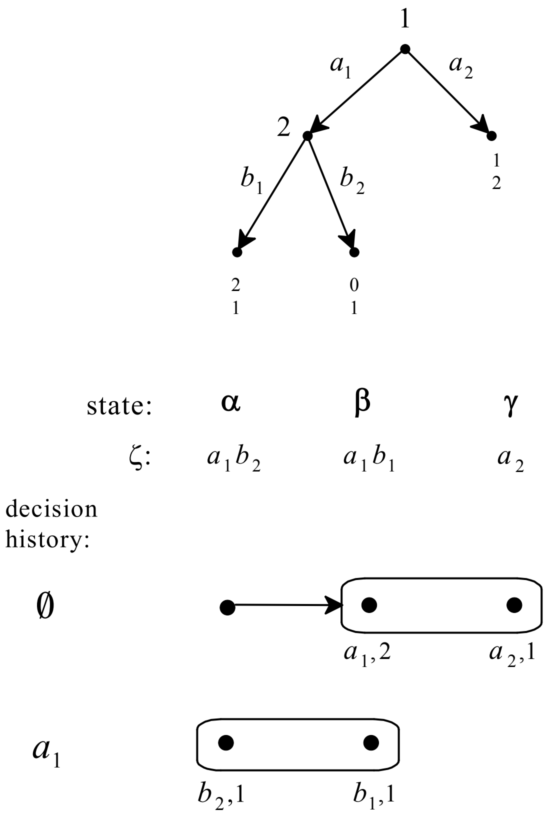

Figure 2: A game and a model of it

must have erroneous beliefs, that is, the relation $\mathcal{B}_h$ is not reflexive at that state. For example, in the model of Figure 1 at history $a_1a_2$ imposing reflexivity of the belief relation of Player 3 would require $\mathcal{B}_{a_1a_2}(\gamma) = \mathcal{B}_{a_1a_2}(\delta) = \mathcal{B}_{a_1a_2}(\epsilon) = \{\gamma, \delta, \epsilon\}$, implying complete uncertainty in the mind of Player 3 as to what will happen if she plays $a_3$.

Thus we need to define correctness of beliefs *locally*, that is, as an event, which may or may not hold at a particular state.

**Definition 4.1.** Let $\mathbf{T}_h \subseteq \Omega$ (where T stands for 'Truth') be the event that the active player at decision history $h$ has correct beliefs:

$$\omega \in \mathbf{T}_h \text{ if and only if } h \prec \zeta(\omega) \text{ and } \omega \in \mathcal{B}_h(\omega).$$

Let $\mathbf{T}$ be the event that, at every reached history, the active player has correct beliefs:

$$\omega \in \mathbf{T} \text{ if and only if, } \forall h \in D \text{ such that } h \prec \zeta(\omega),\ \omega \in \mathbf{T}_h.$$

For example, in the model of Figure 1, $\mathbf{T}_\emptyset = \{\alpha, \beta, \gamma\}$, $\mathbf{T}_{a_1} = \{\beta, \gamma\}$, $\mathbf{T}_{a_1a_2} = \{\gamma, \delta\}$, $\mathbf{T}_{a_1a_2a_3} = \{\delta, \epsilon\}$ and $\mathbf{T} = \{\alpha, \beta, \gamma\}$; in the model of Figure 2, $\mathbf{T}_\emptyset = \{\beta, \gamma\}$, $\mathbf{T}_{a_1} = \{\alpha, \beta\}$ and $\mathbf{T} = \{\beta, \gamma\}$.

The example of Figure 3 shows that rationality and correct beliefs at every reached decision history are not sufficient to guarantee the play of a Nash equilibrium outcome; that is, even if $\omega \in \mathbf{T} \cap \mathbf{R}$, it is not necessarily the case that $\zeta(\omega)$ is a Nash equilibrium outcome. Here we have that $\mathbf{T} = \{\alpha, \beta, \gamma\}$ and $\mathbf{R} = \{\beta, \delta\}$ so that $\mathbf{T} \cap \mathbf{R} = \{\beta\}$,[18] but $\zeta(\beta) = a_1b_1$ which is not a Nash equilibrium outcome.[19] The issue in this case is that Player 1 is uncertain as to what will happen if she plays $a_1$: she does not know whether Player 2 will play $b_1$ or $b_2$; since Player 1 considers it possible that, if she plays $a_1$, Player 2 will play $b_2$ (state $\alpha$) and believes that her choice of $a_2$ would be followed by Player 2 playing $c_1$ (state $\gamma$) and $u_1(a_1b_2) = 2 > u_1(a_2c_1) = 1$, it is rational for her to play $a_1$ (see Definition 3.3). Thus we need to add one more restriction on beliefs, namely that a player is not uncertain as to what will happen if she chooses any particular action.

[18] In this model $\mathbf{T}_\emptyset = \{\alpha, \beta, \gamma\}, \mathbf{T}_{a_1} = \{\alpha, \beta\}, \mathbf{T}_{a_2} = \{\gamma, \delta\}, \mathbf{R}_\emptyset = \Omega, \mathbf{R}_{a_1} = \{\beta\}$ and $\mathbf{R}_{a_2} = \{\delta\}$.

[19] If Player 2's strategy selects choice $b_1$ at decision history $a_1$, then Player 1's best reply is to play $a_2$ rather than $a_1$.

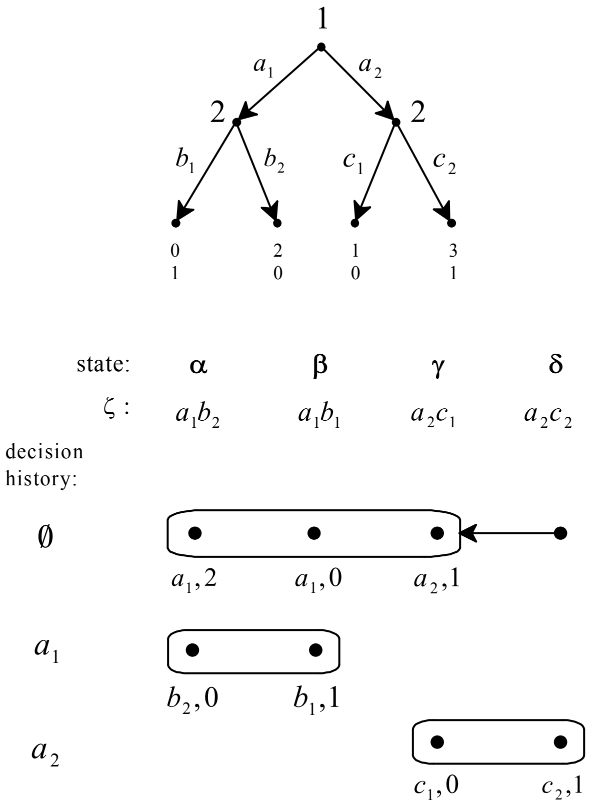

Figure 3: A game and a model of it

**Definition 4.2.** Let $\mathbf{C}_h \subseteq \Omega$ (where $C$ stands for 'Certainty') be the event that the active player at decision history $h$ has no uncertainty about what will happen after any of her choices at $h$:

$$\begin{gathered}\omega \in \mathbf{C}_h \ \text{ if and only if } \ h \prec \zeta(\omega) \ \text{ and, } \ \forall a \in A(h), \forall \omega', \omega'' \in \mathcal{B}_h(\omega), \\ \text{if } \ ha \preceq \zeta(\omega') \ \text{ and } \ ha \preceq \zeta(\omega'') \ \text{ then } \ \zeta(\omega') = \zeta(\omega'').\end{gathered}$$

Let $\mathbf{C}$ be the event that at every reached decision history the active player has no uncertainty about what will happen after any of her choices:

$$\omega \in \mathbf{C} \ \text{ if and only if, } \ \forall h \in D \ \text{ such that } \ h \prec \zeta(\omega), \ \omega \in \mathbf{C}_h.$$

For example, in the model of Figure 3, $\mathbf{C}_\emptyset = \emptyset$,[20] $\mathbf{C}_{a_1} = \{\alpha, \beta\}$ and $\mathbf{C}_{a_2} = \{\gamma, \delta\}$, so that $\mathbf{C} = \emptyset$; in the model of Figure 2, $\mathbf{C}_\emptyset = \{\alpha, \beta, \gamma\}$ and $\mathbf{C}_{a_1} = \{\alpha, \beta\}$ so that $\mathbf{C} = \{\alpha, \beta, \gamma\}$.

**Remark 2.** *Note that the event $\mathbf{C}_h$ only expresses the fact that at history $h$ the active player has no doubt as to what actions future players will take; however her "doubtless" beliefs might be erroneous. In other words, if $\omega \in \mathbf{C}_h$ then at history $h$ player $\iota(h)$ might be certain that after her own action $a$ the subsequent player will play $b$ and yet, as a matter of fact, at state $\omega$ action $a$ is followed (at history $ha$)* not *by action $b$ but by a different action $c$. If, however, $\omega$ belongs to the intersection of events $\mathbf{C}_h$ and $\mathbf{T}_h$ then, at state $\omega$, player $\iota(h)$ has correct beliefs about what will happen after the action she* actually *takes (at $h$ and $\omega$), while there is no way of telling whether or not she is also correct about what would happen after alternative choices at $h$, because the models that we are considering are not reach enough to address that issue (see Remark 6 below).*

Before stating the main result of this section, we need one more definition.

**Definition 4.3.** A game satisfies the no-consecutive-moves condition if no player moves more than once along any given play, that is, if, $\forall h, h' \in D$, $h \prec h'$ implies $\iota(h) \neq \iota(h')$.[21]

The following proposition provides a doxastic characterization of the set of Nash equilibrium *outcomes* (or terminal histories) for games that satisfy the no-consecutive-moves condition.[22] The characterizing condition is that at every

[20] Because, for every $\omega \in \Omega$, $\alpha, \beta \in \mathcal{B}_\emptyset(\omega)$, $a_1 \prec \zeta(\alpha)$, $a_1 \prec \zeta(\beta)$ and $\zeta(\alpha) = a_1 b_2 \neq \zeta(\beta) = a_1 b_1$.

[21] The so-called *agent form* of a game is obtained by treating a player at different decision histories as different players with the same payoff function. Thus the agent form of a game satisfies the no-consecutive-moves condition (but the latter is a weaker condition). Several papers in the literature on the epistemic foundations of backward induction in perfect-information games restrict attention to games in agent form (see, for example, Balkenborg and Winter (1997), Stalnaker (1998)).

[22] At the end of this section we discuss how this restriction can be relaxed.

*reached* history, the active player (1) has correct beliefs, (2) is rational and (3) has no uncertainty about what will happen after any of her choices. This condition is expressed by the event $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$. The proof is given in the Appendix.

**Proposition 1.** *Consider a perfect-information game G that satisfies the no-consecutive-moves condition. Then,*

*(A) If terminal history z is the outcome of a pure-strategy Nash equilibrium of G then there is a model of G and a state $\omega$ in that model such that (1) $\zeta(\omega) = z$ and (2) $\omega \in \mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$.*

*(B) For any model of G and for every state $\omega$ in that model, if $\omega \in \mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$ then there is a pure-strategy Nash equilibrium of G whose corresponding outcome is $\zeta(\omega)$.*

Note that it is only Part *B* of Proposition 1 that requires the restriction to games that satisfy the no-consecutive-moves condition: Part *A* is true for arbitrary perfect-information games. To see why the restriction is needed for Part *B*, consider the game and model of Figure 4, where $\alpha \in \mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$[23] and yet $\zeta(\alpha) = a_2$ which is not a Nash equilibrium outcome: if Player 2's strategy is $b_1$ then Player 1's best reply is either $(a_1, c_1)$ or $(a_1, c_2)$, with corresponding outcome $a_1 b_1$, and if Player 2's strategy is $b_2$ then Player 1's best reply is $(a_1, c_2)$, with corresponding outcome $a_1 b_2 c_2$. The reason why Player 1 is nevertheless rational at state $\alpha$, where she plays $a_2$, is that - in her beliefs - she takes her own choice of $c_1$ at her future decision history $a_1 b_2$ as given, while changing her plan of action at the root from $a_2$ to "first $a_1$ and then $c_2$" would increase her payoff from 1 to 2, making her choice of $a_2$ irrational.

In order to obtain a general version of Proposition 1, that does not require the restriction to games that satisfy the no-consecutive-moves condition, all is needed is a modification of Point 4 of Definition 2.2 where 'action at $h$' is replaced with 'plan of action at $h$'. A plan of action for player $i$ at her decision history $h$ is defined as follows. Let $\{h_1, ..., h_m\}$ be the (possibly empty) set of decision histories of player $i$ that are successors of $h$ (that is, for every $j = 1, ..., m, h \prec h_j$); then a plan of action of player $i$ at $h$ is a pair $(a, \{a_1, ..., a_m\})$ where $a$ is an action at $h$ ($a \in A(h)$) and, for every $j = 1, ..., m$, $a_j$ is an action at $h_j$ ($a_j \in A(h_j)$); if there are no successors of $h$ that are decision histories of player $i$, then a plan of action at $h$ coincides with an action at $h$. The modified Point 4 of Definition 2.2 would require that, for every plan of action of player $i$ at her decision history $h$ there be a state that player $i$ considers possible at $h$ where she "plays" that plan of action. We have opted for the simpler version of Definition 1 because it turns

[23] In this model, $\mathbf{T} = \{\alpha, \beta\}$, $\mathbf{R} = \{\alpha\}$ and $\mathbf{C} = \Omega$, so that $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C} = \{\alpha\}$.

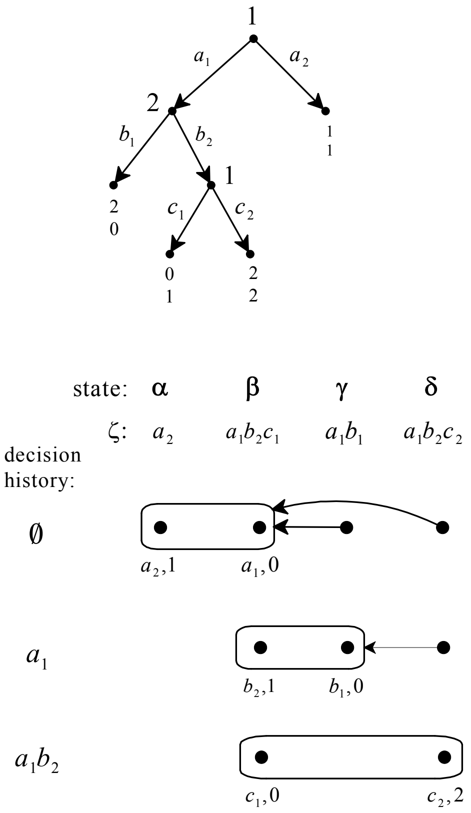

Figure 4: A game and a model of it

out to be sufficient for a characterization of the stronger - and more appealing - notion of backward induction, without requiring the restriction to games that satisfy the no-consecutive-moves condition. The characterization is given in the following section.

# 5 Backward induction

By Proposition 1, the conditions expressed by the event $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$ characterize the notion of Nash equilibrium play in perfect-information games. It follows that for backward induction one needs stronger conditions, since not every Nash equilibrium outcome is a backward-induction outcome. In this section we identify the additional conditions that are required for a characterization of backward induction. Intuitively, rationality is not enough, because it is also necessary for a player to believe that, no matter what action she takes, the next player will act rationally and will believe that future players will act rationally, and so on. To see this, consider the game and model shown in Figure 5, where $\mathbf{T} = \{\alpha, \beta\}$, $\mathbf{R} = \{\beta\}$ and $\mathbf{C} = \Omega$, so that $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C} = \{\beta\}$ and yet $\zeta(\beta) = a_1 b_1$, which is not a backward-induction outcome (the unique backward-induction outcome is $a_1 b_2 c_2$). At state $\beta$ and history $a_1$ Player 2, although rational himself, does not believe that, after his choice of $b_2$, Player 1 will act rationally: he believes that Player 1 would follow with $c_1$ ($\gamma \in \mathcal{B}_a(\beta)$ and $\zeta(\gamma) = a_1 b_2 c_1$). It follows that, since $\beta \in \mathcal{B}_\emptyset(\alpha)$, at state $\alpha$ it is not the case that Player 1 at history $\emptyset$ believes that if she takes action $a_1$ then at history $a_1$ Player 2 will believe that if he takes action $b_2$ then Player 1 will play rationally at history $a_1 b_2$. This is what we need to rule out in order to obtain a characterization of backward induction.

Before proceeding we need to introduce a definition.

**Definition 5.1.** Consider a model of a perfect-information game. Let $\omega \in \Omega$ be a state and let $h = a_1 a_2 ... a_m$ ($m \geq 1$) be a decision history (thus $a_1 \in A(\emptyset)$ and, for every $i = 2, ..., m$, $a_i \in A(a_1 ... a_{i-1})$). We say that *$h$ is reachable from $\omega$* if there exists a sequence $\langle \omega_0, \omega_1, ..., \omega_m \rangle$ in $\Omega$ such that: (1) $\omega_0 = \omega$, (2) for every $i = 1, ..., m$, $a_1 ... a_i \prec \zeta(\omega_i)$, (3) $\omega_1 \in \mathcal{B}_\emptyset(\omega_0)$ and, for every $i = 2, ..., m$, $\omega_i \in \mathcal{B}_{a_1 ... a_{i-1}}(\omega_{i-1})$. We say that any such sequence $\langle \omega_0, \omega_1, ..., \omega_m \rangle$ *leads from $\omega$ to $h$*.

For example, in the model of Figure 1, decision history $a_1 a_2 a_3$ is reachable from $\alpha$ via the sequence $\langle \alpha, \beta, \gamma, \delta \rangle$.[24]

[24] $\beta \in \mathcal{B}_\emptyset(\alpha), a_1 \prec \zeta(\beta) = a_1 d_2, \gamma \in \mathcal{B}_{a_1}(\beta), a_1 a_2 \prec \zeta(\gamma) = a_1 a_2 d_3, \delta \in \mathcal{B}_{a_1 a_2}(\gamma)$ and $a_1 a_2 a_3 \prec \zeta(\delta) = a_1 a_2 a_3 d_4$. Another sequence that leads from $\alpha$ to $a_1 a_2 a_3$ is $\langle \alpha, \gamma, \gamma, \delta \rangle$.

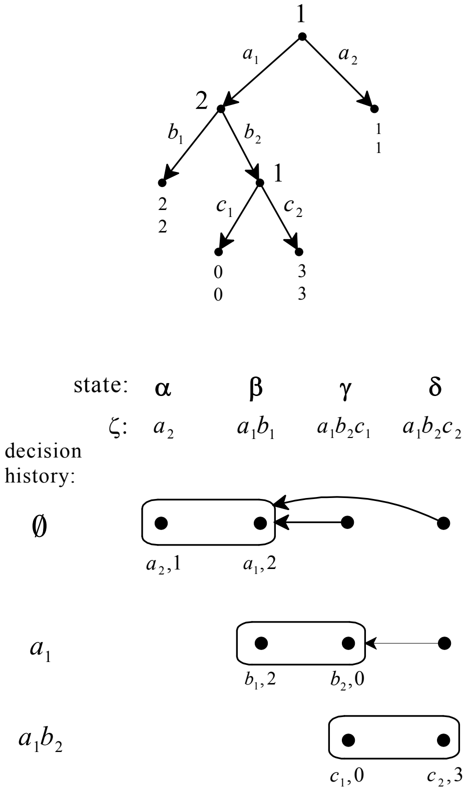

Figure 5: A game and a model of it

**Remark 3.** *Consider a model of a perfect-information game. Let $\omega \in \Omega$ be an arbitrary state and $h \in D$ an arbitrary decision history. Then there is a sequence that leads from $\omega$ to $h$.*[25]

By Proposition 1, the play of Nash equilibrium outcomes is characterized by the conditions expressed by the events $\mathbf{T_h} \cap \mathbf{R_h} \cap \mathbf{C_h}$, for every reached history $h$. We now show that, in order to characterize the play of backward induction outcomes, we need to add "forward iterated belief" in $\mathbf{T_h} \cap \mathbf{R_h} \cap \mathbf{C_h}$.

**Definition 5.2.** Consider a model of a perfect-information game. Let $\mathbf{I}_{TRC}$ (where $I$ stands for 'Iterated') be the following event: $\omega \in \mathbf{I}_{TRC}$ if and only if, for every decision history $h = a_1...a_m$ $(m \geq 1)$, and for every sequence $\langle \omega_0, \omega_1, ..., \omega_m \rangle$ leading from $\omega$ to $h$, $\omega_m \in \mathbf{T}_h \cap \mathbf{R}_h \cap \mathbf{C}_h$.

Definition 5.2 expresses the following condition: the root player (player $\iota(\emptyset)$) believes that if she takes action $a$ and $a$ is a decision history then: (1) the active player at $a$ has correct beliefs, is rational and has no uncertainty, (2) the active player at $a$ believes that if he takes action $b$ and $ab$ is a decision history then the active player at $ab$ has correct beliefs, is rational and has no uncertainty and believes that if she takes action $c$ and $abc$ is a decision history then the active player at $abc$ has correct beliefs, is rational and has no uncertainty, and so forth. All of this can be expressed formally using belief operators, as explained at the end of this section.

So far, characterizations of backward induction have been provided for perfect-information games in generic position (or that satisfy the somewhat weaker condition of "no relevant ties"),[26] which have a unique backward-induction solution. By contrast, the following characterization applies to arbitrary perfect-information games, that is, also those that have multiple backward-induction solutions (note that, unlike Proposition 1, the characterization given in Proposition 2 is *not* restricted to games that satisfy the no-consecutive-moves condition). The proof is given in the Appendix.

[25] Proof. Let $h = a_1...a_m$ $(m \geq 1)$. By Point 1 of Definition 2.2, $\mathcal{B}_\emptyset(\omega) \neq \varnothing$ (since $\emptyset$ is a prefix of every history, in particular of history $\zeta(\omega)$). Hence, since $a_1 \in A(\emptyset)$, by Point 4 of Definition 2.2 there exists an $\omega_1 \in \mathcal{B}_\emptyset(\omega)$ such that $a_1 \preceq \zeta(\omega_1)$. Thus, since $a_2 \in A(a_1)$, by Point 4 of Definition 2.2, there exists an $\omega_2 \in \mathcal{B}_{a_1}(\omega_1)$ such that $a_1 a_2 \preceq \zeta(\omega_2)$, etc.

[26] See, for example, Aumann (1995), Balkenborg and Winter (1997), Ben-Porath (1997), Bonanno (2013), Clausing (2003), Halpern (2001), Perea (2012; 2014), Quesada (2003), Samet (1996; 2013), Stalnaker (1998). A perfect-information game has *no relevant ties* if, $\forall i \in N$, $\forall h \in D_i$, $\forall a, a' \in A(h)$ with $a \neq a'$, $\forall z, z' \in Z$, if $ha$ is a prefix of $z$ and $ha'$ is a prefix of $z'$ then $u_i(z) \neq u_i(z')$. All games in generic position satisfy this condition.

**Proposition 2.** *Consider a perfect-information game G. Then,*

*(A) If terminal history z is the outcome of a backward-induction solution of G then there is a model of G and a state $\omega$ in that model such that (1) $\zeta(\omega) = z$ and (2) $\omega \in (\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC}$.*

*(B) For any model of G and for every state $\omega$ in that model, if $\omega \in (\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC}$ then there is a backward-induction solution of G whose corresponding outcome is $\zeta(\omega)$.*

**Remark 4.** *Note that the event $(\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC}$ is equivalent to the seemingly weaker event $(\mathbf{T}_\emptyset \cap \mathbf{R}_\emptyset \cap \mathbf{C}_\emptyset) \cap \mathbf{I}_{TRC}$.*[27]

**Remark 5.** *In games that have no relevant ties one can dispense with the events $\mathbf{C}_h$, since the no uncertainty condition is a consequence of (loosely speaking) the uniqueness of a rational choice at every decision history.*[28] *Thus for this subclass of games, the event that characterizes the set of backward-induction outcomes is $(\mathbf{T} \cap \mathbf{R}) \cap \mathbf{I}_{TR}$.*[29]

## 5.1 Expressing the event $\mathbf{I}_{TRC}$ using belief operators

We now turn to a discussion on how to interpret the event $\mathbf{I}_{TRC}$ in terms of iterated beliefs, using belief operators.[30] Recall that, given a belief relation

[27] Proof. It is clear that $(\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC} \subseteq (\mathbf{T}_\emptyset \cap \mathbf{R}_\emptyset \cap \mathbf{C}_\emptyset) \cap \mathbf{I}_{TRC}$ since $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C} \subseteq \mathbf{T}_\emptyset \cap \mathbf{R}_\emptyset \cap \mathbf{C}_\emptyset$. To prove the converse, let $\omega \in (\mathbf{T}_\emptyset \cap \mathbf{R}_\emptyset \cap \mathbf{C}_\emptyset) \cap \mathbf{I}_{TRC}$ and let $h = a_1...a_m$ $(m \geq 1)$ be a decision history such that $h \prec \zeta(\omega)$; we need to show that $\omega \in \mathbf{T}_h \cap \mathbf{R}_h \cap \mathbf{C}_h$. Since $\omega \in \mathbf{I}_{TRC}$, it will be sufficient to show that $h$ is reachable from $\omega$ via the constant sequence $\langle \omega_0, \omega_1, ..., \omega_m \rangle$ with $\omega_i = \omega$ for every $i = 0, 1, ..., m$. Point (1) of Definition 5.1 is trivially true and point (2) follows from the hypothesis that $h \prec \zeta(\omega)$ and the fact that, for every $i = 1, ..., m-1$, $a_1...a_i \prec h$. As for Point (3), we have, first of all, that $\omega_1 = \omega \in \mathcal{B}_\emptyset(\omega_\emptyset = \omega)$ because $\omega \in \mathbf{T}_\emptyset$. Thus $a_1$ is reachable from $\omega$ through the sequence $\langle \omega, \omega \rangle$ and hence, since $\omega \in \mathbf{I}_{TRC}$, $\omega \in \mathbf{T}_{a_1}$, that is, $\omega \in \mathcal{B}_{a_1}(\omega)$. It follows that $a_1 a_2$ is reachable from $\omega$ through the sequence $\langle \omega, \omega, \omega \rangle$ and hence, since $\omega \in \mathbf{I}_{TRC}$, $\omega \in \mathbf{T}_{a_1 a_2}$, that is, $\omega \in \mathcal{B}_{a_1 a_2}(\omega)$, and so forth.

[28] The proof is by induction. At a "last" decision node (that is, a decision node followed only by terminal nodes) there is a unique rational choice, since there are no ties. Hence at an immediately preceding node the active player who believes that after each of her choices the corresponding player will play rationally, cannot have uncertainty about the subsequent choices of those future player; hence, since there are no relevant ties, also this player has a unique rational choice. One then extends this argument backwards in the tree by induction.

[29] The event $\mathbf{I}_{TR}$ is defined as in Definition 5.2 but without reference to the events $\mathbf{C}_h$: $\omega \in \mathbf{I}_{\mathbf{TR}}$ if and only if, for every decision history $h = a_1...a_m$ $(m \geq 1)$, and for every sequence $\langle \omega_0, \omega_1, ..., \omega_m \rangle$ leading from $\omega$ to $h$, $\omega_m \in \mathbf{T}_h \cap \mathbf{R}_h$.

[30] The interpretation of the event $\mathbf{I}_{TRC}$ given below in terms of "forward belief in rationality" is conceptually similar to the notion of "forward belief in material rationality" given in (Perea 2007a, Definition 2.7). However, the latter definition is obtained in a class of models where the space of uncertainty is the set of the opponents' strategies, rather than the set of terminal histories

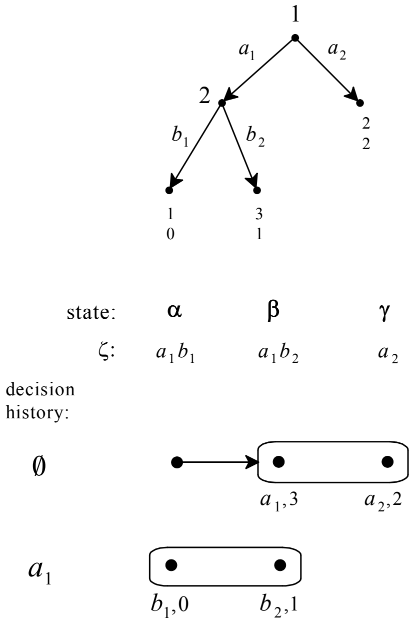

Figure 6: A game and a model of it

$\mathcal{B} \subseteq \Omega \times \Omega$ and an event $E$, we say that at state $\omega$ the individual believes $E$ if $\mathcal{B}(\omega) \subseteq E$ and use this to define a belief operator $\mathbb{B} : 2^\Omega \to 2^\Omega$ as follows: for every $E \subseteq \Omega$, $\mathbb{B}E = \{\omega \in \Omega : \mathcal{B}(\omega) \subseteq E\}$. Consider the game and model of Figure 6. At state $\alpha$ and history $\emptyset$ Player 1 believes that if she plays $a_1$ then Player 2 will play $b_2$, which is a rational choice for Player 2 (indeed, the only rational choice). We would like to express this by writing $\alpha \in \mathbb{B}_\emptyset \mathbf{R}_{a_1}$.[31] However, this is not the case because $\mathbf{R}_{a_1} = \{\beta\}$ and $\mathcal{B}_\emptyset(\alpha) = \{\beta, \gamma\}$ so that $\mathcal{B}_\emptyset(\alpha) \not\subseteq \mathbf{R}_{a_1}$. There are two ways of addressing this issue.

**1.** Using material conditionals. In propositional logic, the material condi-

(furthermore, Perea uses the "type-space" approach rather than the state-space approach followed in this paper). The difference between the two classes of models is discussed in Section 7.2.

[31] That is, at state $\alpha$ and history $\emptyset$, player $\iota(\emptyset) = 1$ believes that at history $a_1$ player $\iota(a_1) = 2$ will act rationally.

tional 'if $p$ then $q$' is true when either $p$ is false or $q$ is true. Correspondingly, the set of states where it is true that 'if event $E$ occurs then event $F$ occurs' is represented by the event $\neg E \cup F$ (where $\neg E$ denotes the complement of $E$). Given a perfect-information game and a model of it, for any decision history $h$ denote by $[h]$ the set of states where $h$ is reached:

$$[h] = \{\omega \in \Omega : h \prec \zeta(\omega)\}.$$

For example, in the model of Figure 6, $[a_1] = \{\alpha, \beta\}$ so that the material conditional 'if Player 1 plays $a_1$ then Player 2 chooses rationally at $a_1$' is represented by the event $\neg[a_1] \cup \mathbf{R}_{a_1} = \{\gamma\} \cup \{\beta\}$ and thus we do have that, at state $\alpha$, Player 1 believes that if she plays $a_1$ then Player 2 will act rationally at history $a_1$: $\alpha \in \mathbb{B}_\emptyset(\neg[a_1] \cup \mathbf{R}_{a_1})$ (since $\mathcal{B}_\emptyset(\alpha) = \{\beta, \gamma\} \subseteq \neg[a_1] \cup \mathbf{R}_{a_1} = \{\beta, \gamma\}$). In the model of Figure 6 the event $\mathbf{I}_{TRC}$ coincides with the event $\mathbb{B}_\emptyset(\neg[a_1] \cup \mathbf{R}_{a_1}) = \{\alpha, \beta, \gamma\}$. Since $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C} = \{\beta\}$, it follows that $(\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC} = \{\beta\}$ and, in accordance to Proposition 2, $\zeta(\beta) = a_1 b_2$ is the backward-induction outcome.

In the model of Figure 7, the event $\mathbf{I}_{TRC}$ is the intersection of the following events:[32]

- $\mathbb{B}_\emptyset\Big(\neg[a_1] \cup (\mathbf{T}_{a_1} \cap \mathbf{R}_{a_1} \cap \mathbf{C}_{a_1})\Big) = \mathbb{B}_\emptyset(\{\beta, \gamma, \delta, \epsilon\}) = \{\gamma, \delta, \epsilon, \eta\}$
- $\mathbb{B}_\emptyset\Big(\neg[a_2] \cup (\mathbf{T}_{a_2} \cap \mathbf{R}_{a_2} \cap \mathbf{C}_{a_2})\Big) = \mathbb{B}_\emptyset(\{\alpha, \beta, \delta, \epsilon, \eta\}) = \Omega$
- $\mathbb{B}_\emptyset \mathbb{B}_{a_1}\Big(\neg[b_1] \cup (\mathbf{T}_{a_1 b_1} \cap \mathbf{R}_{a_1 b_1} \cap \mathbf{C}_{a_1 b_1})\Big) = \mathbb{B}_\emptyset \mathbb{B}_{a_1} \Omega = \mathbb{B}_\emptyset \Omega = \Omega.$[33]

Thus $\mathbf{I}_{TRC} = \{\gamma, \delta, \epsilon, \eta\}$. Since $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C} = \{\beta, \epsilon\}$, $(\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC} = \{\epsilon\}$ and, in accordance to Proposition 2, $\zeta(\epsilon) = a_1 b_2$ is a backward-induction outcome. This game has a second backward-induction outcome, namely $a_2 c_1$; although in this model there is no state $\omega$ such that $\omega \in (\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC}$ and $\zeta(\omega) = a_2 c_1$, by Part *B* of Proposition 2 one can construct a model of this game where such a state exists.

In general, using belief operators and the material conditional, the event $\mathbf{I}_{TRC}$ is the intersection of all the following events, for every decision history $h = a_1 a_2 ... a_m$ $(m \geq 1)$:

[32] In this model, $\mathbf{R}_\emptyset = \{\beta, \epsilon, \eta\}, \mathbf{R}_{a_1} = \{\epsilon\}, \mathbf{R}_{a_2} = \{\beta, \delta\}, \mathbf{R}_{a_1 b_1} = \{\alpha, \eta\}, \mathbf{R} = \{\beta, \epsilon\}, \mathbf{T}_\emptyset = \{\alpha, \beta, \delta, \epsilon\}, \mathbf{T}_{a_1} = \{\epsilon, \eta\}, \mathbf{T}_{a_2} = \{\beta, \gamma, \delta\}, \mathbf{T}_{a_1 b_1} = \{\alpha, \eta\}, \mathbf{T} = \{\beta, \delta, \epsilon\}, \mathbf{C}_\emptyset = \Omega, \mathbf{C}_{a_1} = \{\alpha, \epsilon, \eta\}, \mathbf{C}_{a_2} = \{\beta, \gamma, \delta\}, \mathbf{C}_{a_1 b_1} = \{\alpha, \eta\}, \mathbf{C} = \Omega, [a_1] = \{\alpha, \epsilon, \eta\}, [a_2] = \{\beta, \gamma, \delta\}$ and $[b_1] = \{\alpha, \eta\}$.

[33] Note that, if state $\omega$ and decision history $h$ are such that $h$ is not reached at $\omega$ (that is, $h \nprec \zeta(\omega)$), then, by Definition 2.2, $\mathcal{B}_h(\omega) = \varnothing$ and therefore $\mathcal{B}_h(\omega) \subseteq E$, for every event $E$, that is, $\omega \in \mathbb{B}_h E$. For example, in the model of Figure 7, $\mathbb{B}_{a_1}(\{\alpha, \beta\}) = \{\beta, \gamma, \delta\}$, since, for every $\omega \in \{\beta, \gamma, \delta\}$, $\mathcal{B}_{a_1}(\omega) = \varnothing$.

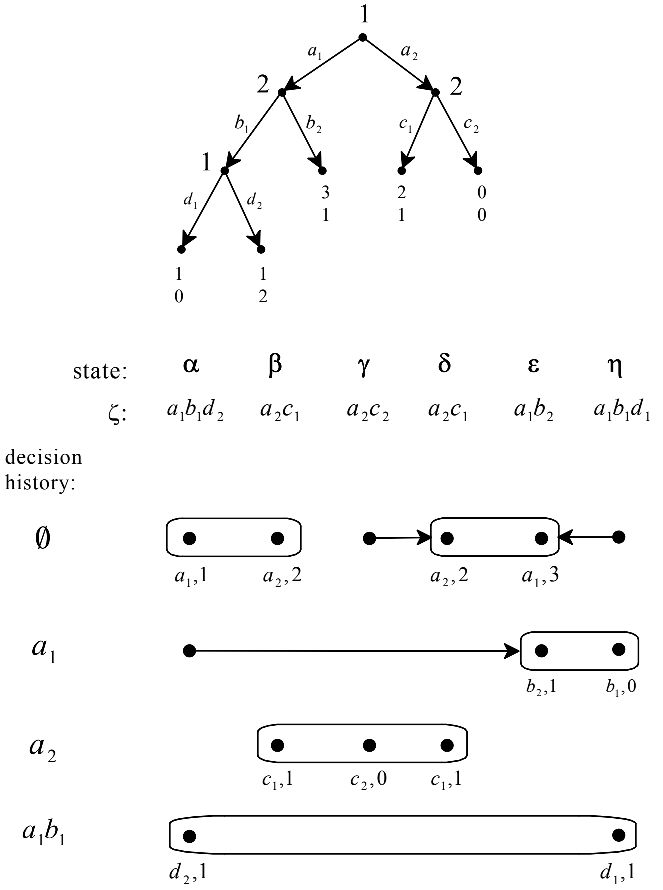

Figure 7: A game and a model of it

- $\mathbb{B}_\emptyset\Big(\neg[a_1] \cup (\mathbf{T}_{a_1} \cap \mathbf{R}_{a_1} \cap \mathbf{C}_{a_1})\Big)$
- $\mathbb{B}_\emptyset\mathbb{B}_{a_1}\Big(\neg[a_2] \cup (\mathbf{T}_{a_1a_2} \cap \mathbf{R}_{a_1a_2} \cap \mathbf{C}_{a_1a_2})\Big)$
- ...
- $\mathbb{B}_\emptyset\mathbb{B}_{a_1}\mathbb{B}_{a_1a_2}...\mathbb{B}_{a_1a_2...a_{m-1}}\Big(\neg[a_m] \cup (\mathbf{T}_{a_1a_2...a_m} \cap \mathbf{R}_{a_1a_2...a_m} \cap \mathbf{C}_{a_1a_2...a_m})\Big)$.

**2.** The alternative (and essentially equivalent) approach is to extend the definitions of $\mathbf{T}_h, \mathbf{R}_h$ and $\mathbf{C}_h$ by adding to those events all the states at which history $h$ is not reached. In other words, define the following events:

$$\hat{\mathbf{T}}_h = \neg[h] \cup \mathbf{T}_h, \quad \hat{\mathbf{R}}_h = \neg[h] \cup \mathbf{R}_h, \quad \hat{\mathbf{C}}_h = \neg[h] \cup \mathbf{C}_h \tag{1}$$

Using this approach the event $\mathbf{I}_{TRC}$ is the intersection of all the following events, for every decision history $h = a_1a_2...a_m$ $(m \geq 1)$:

- $\mathbb{B}_\emptyset(\hat{\mathbf{T}}_{a_1} \cap \hat{\mathbf{R}}_{a_1} \cap \hat{\mathbf{C}}_{a_1})$
- $\mathbb{B}_\emptyset\mathbb{B}_{a_1}(\hat{\mathbf{T}}_{a_1a_2} \cap \hat{\mathbf{R}}_{a_1a_2} \cap \hat{\mathbf{C}}_{a_1a_2})$
- ...
- $\mathbb{B}_\emptyset\mathbb{B}_{a_1}\mathbb{B}_{a_1a_2}...\mathbb{B}_{a_1a_2...a_{m-1}}(\hat{\mathbf{T}}_{a_1a_2...a_m} \cap \hat{\mathbf{R}}_{a_1a_2...a_m} \cap \hat{\mathbf{C}}_{a_1a_2...a_m})$.

# 6 Related literature

To the best of our knowledge, with one exception (Ben-Porath (1997))[34] existing characterizations of Nash equilibrium have been restricted to strategic-form games (Aumann and Brandenburger (1995), Bach and Tsakas (2014), Barelli (2009), Perea (2007b), Polak (1999)). Proposition 1 provides a characterization of Nash equilibrium *outcomes* in perfect-information games in terms of three conditions at reached decision histories: (1) correct beliefs (event **T**), (2) rationality (event **R**) and (3) no uncertainty (event **C**). Before comparing our characterization to Ben Porath's characterization we point out the limited sense in which beliefs are postulated to be "correct", that is, the limited sense in which event **T** captures the notion of correct belief.

[34] It should also be noted that, for perfect-information games with no relevant ties, Battigalli et al. (2013) shows that in every type structure there is a unique play consistent with common strong belief of material rationality and that play is a Nash equilibrium play.

**Remark 6.** *Given a decision history $h$, the event $\mathbf{T}_h$ (Definition 4.1) captures the notion of correct beliefs: if $\omega \in \mathbf{T}_h$, that is, if $\omega \in \mathcal{B}_h(\omega)$ (for $h$ such that $h \prec \zeta(\omega)$), then the active player at history $h$ does not have false beliefs. Does this mean that she is correct in her beliefs about what will happen if she takes* any *of her actions at $h$? For the action she* actually *takes at $h$ (that is, for that action $a \in A(h)$ such that $ha \preceq \zeta(\omega)$) her beliefs are indeed correct, but for any other action $a' \in A(h) \setminus \{a\}$ the models that we are using are not sufficiently rich to answer the question, since such an answer would involve the evaluation of a counterfactual, as explained below.*

Consider, for example, state $\delta$ in the model of Figure 7, where at decision history $\emptyset$ Player 1 believes that if she takes action $a_1$ then Player 2 will follow with $b_2$ (state $\epsilon$) and if she takes action $a_2$ then Player 2 will follow with $c_1$ (state $\delta$). The latter belief is correct at state $\delta$, where - as a matter of fact - Player 2 plays $c_1$, but is the former belief also correct, that is, is it true - *at state* $\delta$ - that if Player 1 played $a_1$ then Player 2 would play $b_2$? From the point of view of state $\delta$, the proposition 'if Player 1 plays $a_1$ then Player 2 plays $b_2$' is a counterfactual proposition, that is, one that has a false antecedent. The standard theory of counterfactuals (Lewis (1973), Stalnaker (1968)) requires that we identify a state $\delta'$ which is such that (1) it is "most similar" to state $\delta$ and (2) it is true at $\delta'$ that Player 1 plays $a_1$ (that is, $ha \prec \zeta(\delta')$); one then declares the counterfactual true at state $\delta$ if - at this alternative state $\delta'$ - Player 2 actually plays $b_2$.[35] There is no reason why, in the model of Figure 7, one should take $\epsilon$ to be the closest or most similar state to $\delta$ where Player 1 plays $a_1$; indeed Player 1 might have false beliefs, at state $\delta$, about the consequences of taking action $a_1$. We could enrich our models by adding a counterfactual selection function, but it is not clear what value there would be in such an extension: from the point of view of evaluating the rationality of an action what matters is the player's belief, even if such belief is erroneous.[36]

Theorem 2 in Ben-Porath (1997) states that, if the following conditions hold at a state, then the associated outcome is a Nash equilibrium outcome: (1) there is Common Certainty of rationality, (2) there is Common Certainty that each player assigns positive probability to the true profile of strategies and beliefs of the other players and (3) there is Common Certainty of the support of the beliefs

[35] Stalnaker (Stalnaker (1968)) postulates a "selection function" $f : \Omega \times 2^{\Omega} \to \Omega$ that associates with every state $\omega$ and event $E$ a unique state $f(\omega, E) \in E$, while Lewis (Lewis (1973)) postulates a selection function $F : \Omega \times 2^{\Omega} \to 2^{\Omega}$ that associates with every state $\omega$ and event $E$ a *set* of states $F(\omega, E) \subseteq E$. Stalnaker declares the proposition 'if $E$ then $G$' true at $\omega$ if and only if $f(\omega, E) \in G$, while Lewis requires that $F(\omega, E) \subseteq G$.

[36] For an extensive discussion of this issue see Bonanno (2015).

of each player.[37] A player is certain of an event $A$ if she assigns probability 1 to $A$; there is Common Certainty of $A$ if event $A$ occurred, each player is certain of $A$, each player is certain that every other player is certain of $A$, and so forth. The following are the main differences between our characterization of Nash equilibrium outcomes and Ben Porath's characterization:

- Ben Porath restricts attention to generic games.
- Ben Porath uses a stronger notion of rationality, namely expected utility maximization.
- The models considered by Ben Porath are not behavioral models but strategy-based models: a state specifies a full strategy for each player, rather than just the actions actually taken.[38]
- Ben Porath's characterization is in terms of *common belief*: beliefs about beliefs about beliefs ..., while the characterization we provide is in terms of "facts" and does not require any form of beliefs about beliefs (in our models iterated beliefs are required for backward induction outcomes, not for Nash equilibrium outcomes).[39]

[37] As the author notes, Theorem 2 does not provide a full characterization of Nash equilibrium outcomes as there are Nash equilibria that are inconsistent with extensive-form rationality. However, if only normal-form rationality is assumed, that is, if one assumes that a player optimizes only with respect to her initial beliefs (and not necessarily at every node), then the conditions of Theorem 2 provide a full characterization of Nash equilibrium outcomes.

[38] Furthermore, Ben Porath uses the "type space" approach where a state is identified with an $n$-tuple of types, one for each player ($n$ is the number of players); the type of a player specifies his strategy as well as a belief function that assigns, for every node in the tree, a probabilistic belief over the set of profiles of types of the other players. Each player is assumed to know his own type; in particular, each player knows his own strategy.

[39] Epistemic characterizations of Nash equilibrium in strategic-form games have not relied on the condition of common belief of rationality. For example, in their seminal paper Aumann and Brandenburger Aumann and Brandenburger (1995) showed that, in games with more than two players, if there exists a common prior then mutual belief in rationality and payoffs as well as common belief in each player's conjecture about the opponents' strategies imply Nash equilibrium. However, Polak Polak (1999) later showed that in complete-information games, Aumann and Brandenburger's conditions actually do imply common belief in rationality. More recently, Barelli Barelli (2009) generalized Aumann and Brandenburger's result by substituting the common prior assumption with the weaker property of action-consistency, and common belief in conjectures with a weaker condition stating that conjectures are constant in the support of the action-consistent distribution. Thus, he provided sufficient epistemic conditions for Nash equilibrium without requiring common belief in rationality. Later, Bach and Tsakas Bach and Tsakas (2014) obtained a further generalization by introducing even weaker epistemic conditions for Nash equilibrium than those in Barelli (2009): their characterization of Nash equilibrium is based on introducing pairwise

The characterization of backward-induction outcomes provided in Proposition 2 is in terms of a, by now well-understood,[40] condition, namely that players should believe that future players will be rational and will believe the same about players who will choose after them and so forth. Since the literature on the epistemic foundations of backward induction has been thoroughly reviewed by Perea (Perea (2007a),(Perea 2012, p.463)) and Brandenburger (Brandenburger (2007)) it is unnecessary to go into the details of each contribution. The distinguishing features of our approach were listed in Section 1.

As noted earlier, the purely behavioral point of view that we have adopted (consisting in associating with every state a play of the game rather than a strategy profile) was first introduced by Samet (Samet (1996)). The other papers that take a purely behavioral point of view are based on a specification of each player's initial beliefs as well as her disposition to revise those beliefs in response to information that she might receive during the play of the game; this is done either probabilistically using conditional probability systems (Battigalli et al. (2013)) or by means of qualitative belief revision structures (Baltag et al. (2009), Stalnaker (1996; 1998)). Those models impose the constraint that, if at a state player $i$ chooses strategy $s_i$, then she knows that she uses strategy $s_i$ (that is, at every state that she considers possible, she uses strategy $s_i$); thus subjective counterfactuals (or dispositional belief revision) are needed in order to represent a player's beliefs about the consequences of choosing a strategy different from $s_i$.

The closest paper to the present one is Bonanno (2013), which also uses behavioral models and the weak notion of rationality postulated in this paper. The main differences are: (1) Bonanno (2013) only deals with backward induction and not Nash equilibrium, (2) the characterization of backward induction provided in Bonanno (2013) is obtained only for games without relevant ties and thus does not cover games with multiple backward-induction solutions[41] and (3) the models considered in Bonanno (2013) are explicitly dynamic models, while in the models considered in this paper time plays no explicit role.

epistemic conditions imposed only on *some* pairs of players (contrary to the characterizations in Aumann and Brandenburger (1995) and Barelli (2009), which correspond to pairwise epistemic conditions imposed on *all* pairs of players). Not only do these conditions not imply common belief in rationality but they do not even imply mutual belief in rationality.

[40] See Balkenborg and Winter (1997), Baltag et al. (2009), Clausing (2003; 2004), Feinberg (2005), Perea (2014), Stalnaker (1998).

[41] The conditions for backward induction provided in Bonanno (2013) are conceptually the same as those expressed by the event $(\mathbf{T} \cap \mathbf{R}) \cap \mathbf{I}_{TR}$ (see Remark 5).

# 7 Discussion

## 7.1 On the relationship between backward induction and subgame-perfect equilibrium

An anonymous referee raised the issue whether Proposition 2 should be viewed as a characterization of the notion of subgame-perfect equilibrium (SPE) or of the notion of backward induction solution (BIS). Are the two notions different? This is a subtle point, which hinges on the meaning of the expression "backward induction solution". If by BIS one means "the set of choices selected by the backward-induction algorithm", then the two notions of SPE and BIS coincide in finite games with perfect information (with or without relevant ties): a pure-strategy profile $s$ is a SPE if and only if $s$ is a possible output of the backward-induction algorithm. By giving an alternative interpretation to the notion of "backward induction solution", one can draw a distinction (in perfect-information games *with* relevant ties) between SPE and BIS, but this requires using a different class of models from the one considered in this paper: one needs to consider models where states are interpreted in terms of *strategy profiles* (rather than in terms of terminal histories). For example, in the game of Figure 2, the set of pure-strategy subgame-perfect equilibria is $\{(a_1, b_1), (a_2, b_2)\}$ while one could consider the set of "backward induction solutions" to be the larger set $\{(a_1, b_1), (a_2, b_2), (a_2, b_1)\}$, that is, one can also view $(a_2, b_1)$ as a BIS in the following sense. One can construct a model where the strategy profile associated with a state $\omega$ is $(a_2, b_1)$ and, at $\omega$, both players are rational because (1) Player 1 knows that his strategy is $a_2$ and, on the supposition that he plays $a_1$, believes that Player 2 would play $b_2$ – which is a rational choice for Player 2 – and (2) if Player 1 were to play $a_1$ then Player 2 would play $b_1$ – which would be a rational choice for her. Thus, in such a model, at state $\omega$, Player 1 would have erroneous beliefs about what would happen if he played $a_1$ (instead of the chosen action $a_2$). On the other hand, in the approach proposed in this paper one cannot distinguish between the strategy profile $(a_2, b_2)$ (which is a SPE) and the strategy profile $(a_2, b_1)$ (which is not even a Nash equilibrium), since the associated outcome (namely, $a_2$) is the same, and states are only described in terms of outcomes, not in terms of strategies. Indeed, as pointed out in Remark 6, in the models used in this paper at a state where Player 1 chooses action $a_2$ one cannot even assess whether Player 1's beliefs about what would happen if she played $a_1$ are correct or not. Thus the question whether Proposition 2 should be viewed as a characterization of the notion of subgame-perfect equilibrium or of the notion of backward induction solution (assuming that one is adopting a broad interpretation of the latter) seems to lie, conceptually, outside the scope of the framework adopted in this paper.

Several contributions in the literature have explored the relationship between the notion of SPE and "backward-induction-like" procedures in general extensive-form games (that is, games with possibly imperfect information). Kaminski (Kaminski (2009)) defines a backward-induction equilibrium (BIE) as a strategy profile that survives "backward pruning" and proves that in a large class of extensive-form games a strategy profile is a BIE if and only if it is a SPE, thus extending the equivalence beyond perfect-information games. Penta (Penta (2009)) proposes an alternative extension of the notion of backward induction to extensive games with imperfect information: his "backward rationalizability procedure" iteratively eliminates strategies and conditional belief vectors starting from the end of the game and proceeding backwards towards the root. Perea Perea (2014) provides an epistemic characterization of this procedure in terms of the notion of "common belief in future rationality".[42] He also introduces a new algorithm, the "backward dominance procedure", which differs from Penta's procedure in that it operates only on strategies (rather than strategies and conditional belief vectors)[43] and shows that the strategies that survive the backward dominance procedure are exactly the strategies that can be chosen under common belief in future rationality if one does not impose (common belief in) Bayesian updating. The generalized notions of backward induction proposed by Penta and Perea are weaker than SPE.[44] Both authors rely on the approach commonly used in the literature where the underlying space of uncertainty is the set of the opponents' strategies, while in our approach the underlying space of uncertainty is the set of terminal histories.

[42] Defined as follows: players are rational and always believe in their opponents' present and future rationality and believe that every opponent always believes in his opponents' present and future rationality and that every opponent always believes that every other player always believes in his opponents' present and future rationality, and so on.

[43] In the first round the algorithm eliminates, at every information set of player $i$, strategies of player $i$ himself that are strictly dominated at present and future information sets, as well as strategies of players other than $i$ that are strictly dominated at present and future information sets. In every further round $k$ those strategies are eliminated that are strictly dominated at a present or future information set $I_i$ of player $i$, given the opponents' strategies that have survived up to round $k$ at that information set $I_i$. The strategies that eventually survive the elimination process constitute the output of the backward dominance procedure.

[44] Perea (Perea (2014)) also suggests that, in general extensive-form games, the two notions of common belief in future rationality and sequential equilibrium reflect the difference between BIS and SPE.

## 7.2 On the space of uncertainty: strategy profiles versus terminal histories

There is a large literature on the issue of whether common belief of rationality implies backward induction in perfect-information games (Artemov (2010), Aumann (1995; 1996; 1998), Bach and Heilmann (2011), Balkenborg and Winter (1997), Baltag et al. (2009), Battigalli et al. (2013), Ben-Porath (1997), Binmore (1996; 1997), Clausing (2003; 2004), Halpern (2001), Quesada (2003), Samet (2013), Stalnaker (1998)). Most of this literature has focused on games without relevant ties and has employed models of games where the underlying space of uncertainty is the set of the opponents' strategies. In contrast, in our approach the underlying space of uncertainty is the set of terminal histories. An anonymous reviewer suggested that, by not modeling hypothetical or counterfactual beliefs, one might give up the ability to model the reasoning of the players in dynamic games. The lively debate on the relationship between (common belief of) rationality and backward induction (most notably, Aumann (1995; 1996), Binmore (1996; 1997), Halpern (2001), Stalnaker (1998)) has centered on the issue of what is the correct way of modeling hypothetical or counterfactual beliefs about players' rationality. For example, Stalnaker ((Stalnaker 1998, pp.45-46)) claims that Aumann (Aumann (1995)) "equivocates between epistemic and causal 'if's" and that "a strong and implausible belief revision policy has been implicitly built into [Aumann's] definition of substantive rationality". Referring to a "centipede-like"game (that is, a game with a structure similar to the game shown in Figure 1 above), he illustrates this point as follows ((Stalnaker 1998, p.48))

> Bob has the following initial belief: Alice would choose $A_2$ on her second move if she had a second move. This is a causal 'if' – an 'if' used to express Bob's opinion about Alice's disposition to act in a situation that they both know will not arise. Bob knows that since Alice is rational, if she somehow found herself at the second node, she would choose $A_2$. But to ask what Bob would believe about Alice if he learned that he was wrong about her first choice is to ask a completely different question – this 'if' is epistemic; it concerns Bob's belief revision policies, and not Alice's disposition to be rational. No assumption about Alice's substantive rationality, or about Bob's knowledge of her substantive rationality, can imply that Bob should be disposed to maintain his belief that she will act rationally on her second move even were he to learn that she acted

> irrationally on her first.

While the traditional models (where the underlying space of uncertainty is the set of the opponents' strategies) allow one to address a rich set of issues – in particular, the issues mentioned above concerning belief revision – they also raise the question of what it means to interpret a state in terms of a strategy profile. As Halpern ((Halpern 2001, p.433)) points out, in this type of models "one possible culprit for the confusion in the literature regarding what is required to force the backward induction solution in games of perfect information is the notion of a strategy". For example, consider the game of Figure 1 above and a state where Player 1's strategy is $(d_1, a_4)$:

> According to strategy $(d_1, a_4)$, Player 1 plays $a_4$ at history $a_1a_2a_3$. But $a_1a_2a_3$ is a history that cannot be reached if Player 1 uses the strategy $(d_1, a_4)$, because according to this strategy, Player 1 plays $d_1$ at the root. The standard reading of the strategy $(d_1, a_4)$ is that "if history $a_1a_2a_3$ is reached, then Player 1 plays $a_4$". But this reading leaves a number of questions unanswered. How Player 1 plays (if she is rational) depends on her beliefs. Should this be read as "no matter what Player 1's beliefs are, if history $a_1a_2a_3$ is reached, then Player 1 will play $a_4$"? Or perhaps it should be that "given her current beliefs (regarding, for example, what moves the other players will make), if history $a_1a_2a_3$ is reached, then Player 1 will play $a_4$". Or perhaps it should be "in the state closest to the current state where history $a_1a_2a_3$ is actually reached, Player 1 plays $a_4$". ((Halpern 2001, p.434), paraphrased to match the example of Figure 1).

Halpern adopts the last interpretation and shows that one can make sense of Aumann's (Aumann (1995)) and Stalnaker's (Stalnaker (1998)) opposite claims about the implications of common belief of rationality for backward induction,[45] by explicitly modeling the counterfactuals that are implicit in the strategies and by varying the interpretation of those counterfactuals.

The approach put forward in this paper – which dispenses with strategies and interprets states in terms of plays (or terminal histories) – is certainly not rich enough to provide an epistemic foundation for the backward-induction *strategies*, but it does provide an epistemic foundation for the backward-induction

[45] Aumann's claim is that common knowledge of substantive rationality implies the backward induction solution in perfect-information games without relevant ties, while Stalnaker maintains that it does not. Roughly speaking, a player is substantively rational if, for every history $h$ of hers, if the play of the game were to reach $h$, then she would be rational at $h$.

*play*. A consequence of this is that the subtle issues, discussed in the literature, pertaining to belief revision are not conceptually necessary for an understanding of what is needed to obtain the backward-induction *outcome*.[46] The difference between the two approaches ("states interpreted in terms of strategies" versus "states interpreted in terms of outcomes") reflects a different philosophy concerning the nature of theoretical predictions in game theory. In the strategy-based approach the prediction is in terms not only of what play will be observed, but also in terms of a set of counterfactuals about what the various players would do in circumstances that ought not to arise given the predicted outcome. In the outcome-based approach what a player would actually do at an unreached history is left unspecified; however, *the beliefs of the active players along the actual play of the game about the possible consequences of alternative moves are explicitly modeled and provide the reasoning behind the selection of a specific move*.

## 7.3 On the notion of pre-choice belief

While the standard approach in the literature is to model a player's beliefs *after* she has made her choice (and thus knows what that choice is), we have chosen to model beliefs at the pre-choice, or deliberation, stage. This does *not* mean that we are representing the beliefs of the players before the game is played; on the contrary, beliefs are modeled as beliefs *during the play of the game at decision nodes that are actually reached*. According to this approach, when it is her turn to move, a player considers the possible consequences of all her actions, without pre-judging her subsequent decision; in other words, the beliefs of the active player at a state $\omega$ and decision history $h$ (that is reached at $\omega$) are truly open to the possibility of taking any of the actions available at $h$. This reflects the view, expressed by several authors (Gilboa (1999), Ginet (1962), Goldman (1970), Ledwig (2005), Levi (1986; 1997), Shackle (1958), Spohn (1977; 1999), that it is the essence of deliberation that one cannot reason towards a choice if one already knows what that choice will be. For example, Gilboa writes:

> "[. . . ] we are generally happier with a model in which one cannot be said to have beliefs about (let alone knowledge of) one's own choice *while making this choice*. [. . . ] One may legitimately ask: Can you truly claim you have no beliefs about your own future choices? Can you honestly contend you do not believe - or even

[46] It is worth noting that, as pointed out by Samet ((Samet 2013, p.194), while Aumann (Aumann (1995)) states the weaker claim that common knowledge of substantive rationality implies the backward-induction play, he actually proves that it implies the backward-induction *strategies*.

> know - that you will not choose to jump out of the window? [. . . ] The answer to these questions is probably a resounding "No". But the emphasis should be on timing: when one considers one's choice tomorrow, one may indeed be quite sure that one will not decide to jump out of the window. However, a future decision should actually be viewed as a decision by a different "agent" of the same decision maker. [. . . ] *It is only at the time of choice, within an "atom of decision", that we wish to preclude beliefs about it*." ((Gilboa 1999, pp. 171-172), second emphasis added.)]

An implication of this point of view is that, since – at the time of deliberation – the agent does not know what choice she is going to make, she cannot know that her forthcoming choice is rational. For example, at a state $\omega$ that belongs to the event that characterizes backward induction (event $\mathbf{I}_{TRC}$: see Definition 5.2), at a reached history $h$ the active player does not know that she is rational (at $h$), even though she believes that every future player (and possibly herself at histories that are successors of $h$) is rational. This issue has been discussed extensively in the philosophical literature (see, for example, Levi (1986; 1997), Spohn (1977; 1999)), where it is argued that no inconsistency is involved in this approach (as pointed out in the above quote from Gilboa (1999))).

As pointed out in Section 7.2, the advantage of modeling beliefs as pre-choice beliefs is that one can obtain a "conceptually lighter" characterization of backward induction that does not require the use of (objective or subjective) counterfactuals.

# A Proofs

Before proving Proposition 1 we introduce some notation and a definition.

Let $G$ be a perfect-information game and $\sigma$ a pure-strategy profile of $G$. Let $f_\sigma : H \to Z$ (recall that $H$ is the set of histories and $Z$ is the set of terminal histories) be defined as follows: if $z \in Z$ then $f_\sigma(z) = z$ and if $h \in D$ (recall that $D$ is the set of decision histories) then $f_\sigma(h)$ is the terminal history reached from $h$ by following the choices prescribed by $\sigma$.

**Definition A.1.** Let $G$ be a perfect-information game and $\sigma$ a pure-strategy profile of $G$. The *model of G generated by* $\sigma$ is the following model:

- $\Omega = Z$.

- $\zeta : Z \to Z$ is the identity function: $\zeta(z) = z, \forall z \in Z$.

- For every $h \in D$, $\mathcal{B}_h \subseteq Z \times Z$ is defined as follows: $\mathcal{B}_h(z) \neq \varnothing$ if and only if $h \prec z$ and $z' \in \mathcal{B}_h(z)$ if and only if $z' = f_\sigma(ha)$ for some $a \in A(h)$ (recall that $A(h)$ is the set of actions available at $h$). That is, the active player at decision history $h$ believes that if she takes action $a$ then the outcome will be the terminal history reached from $ha$ by $\sigma$.

Figure 8 shows an extensive form with perfect information and the model generated by the strategy profile $\sigma = (a_1, b_1, c_1, d_1)$ ($\sigma$ is highlighted by double edges).

**Remark 7.** *Let $G$ be a perfect-information game and $\mathcal{M}$ the model generated by a pure-strategy profile $\sigma$ of $G$. Then the no-uncertainty condition (Definition 4.2) is satisfied at every state, that is, $\mathbf{C} = Z$. Furthermore, if $z^*$ is the play generated by $\sigma$ (that is, $z^* = f_\sigma(\emptyset)$), then $z^* \in \mathcal{B}_h(z^*)$ for all $h \in D$ such that $h \prec z^*$; that is, $z^* \in \mathbf{T}$.*

### Proof of Proposition 1

*Proof.* **(A)** Fix a perfect-information game $G$ (not necessarily one that satisfies the no-consecutive-moves condition) and let $\sigma$ be a pure-strategy Nash equilibrium of $G$. If $h$ is a decision history, to simplify the notation we shall write $\sigma(h)$ instead of $\sigma_{\iota(h)}(h)$ to denote the choice selected by $\sigma$ at $h$. Consider the model generated by $\sigma$ (Definition A.1). Let $z^*$ be the play generated by $\sigma$, that is, $z^* = f_\sigma(\emptyset)$. By Remark 7, $z^* \in \mathbf{C} \cap \mathbf{T}$. Thus it only remains to show that $z^* \in \mathbf{R}$, that is, that $z^* \in \mathbf{R}_h$, for all $h \in D$ such that $h \prec z^*$. Fix an arbitrary $h \in D$ such that $h \prec z^*$ and let $a$ be the action at $h$ such that $ha \preceq z^*$, that is, $\sigma(h) = a$; then $f_\sigma(ha) = f_\sigma(\emptyset) = z^*$. Suppose that $z^* \notin \mathbf{R}_h$. Then there is an action $b \in A(h) \setminus \{a\}$ that guarantees a higher utility to player $\iota(h)$, that is, if $z' \in \mathcal{B}_h(z^*)$ is such that $hb \preceq z'$, then $u_{\iota(h)}(z') > u_{\iota(h)}(z^*)$. By Definition A.1, $z' = f_\sigma(hb)$ and thus $u_{\iota(h)}(f_\sigma(hb)) > u_{\iota(h)}(f_\sigma(ha))$ so that by unilaterally changing his strategy at $h$ from $a$ to $b$ (and leaving the rest of his strategy unchanged), player $\iota(h)$ can increase his payoff, contradicting the assumption that $\sigma$ is a Nash equilibrium.

(B) Let $G$ be a perfect-information game that satisfies the no-consecutive-moves condition (Definition 4.3) and consider a model of it where there is a state $\alpha$ such that $\alpha \in \mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$. We need to construct a pure-strategy Nash equilibrium $\sigma$ of $G$ such that $f_\sigma(\emptyset) = \zeta(\alpha)$.

STEP 1. For every $h \in D$ such that $h \prec \zeta(\alpha)$, let $\sigma(h) = a$ where $a \in A(h)$ is the action at $h$ such that $ha \preceq \zeta(\alpha)$.

STEP 2. Fix an arbitrary $h \in D$ such that $h \prec \zeta(\alpha)$ and an arbitrary $b \in A(h)$ such that $b \neq \sigma(h)$ ($\sigma(h)$ was defined in Step 1). Since $\alpha \in \mathbf{C}$, for every $\omega, \omega' \in \mathcal{B}_h(\alpha)$

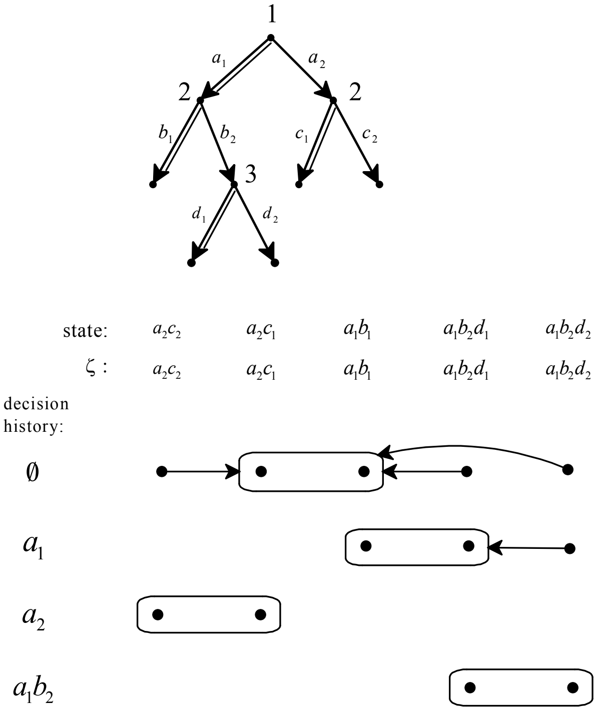

Figure 8: A game and the model generated by the strategy profile $\sigma = (a_1, b_1, c_1, d_1)$

such that $hb \preceq \zeta(\omega)$ and $hb \preceq \zeta(\omega')$, $\zeta(\omega) = \zeta(\omega')$. Select an arbitrary $\omega \in \mathcal{B}_h(\alpha)$ such that $hb \preceq \zeta(\omega)$ and define, for every $h' \in D$ such that $hb \preceq h' \prec \zeta(\omega)$, $\sigma(h') = c$ where $c \in A(h')$ is the action at $h'$ such that $h'c \preceq \zeta(\omega)$.

So far we have defined the choices prescribed by $\sigma$ along the play to $\zeta(\alpha)$ and for paths at one-step deviations from this play. This is illustrated in Figure 9, where $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C} = \{\delta\}$.[47] Focusing on state $\delta$, the above two steps yield the following partial strategy profile (which is highlighted by double edges). By Step 1, $\sigma(\emptyset) = a_1, \sigma(a_1) = b_2$ and, by Step 2, $\sigma(a_2) = d_1, \sigma(a_1b_1) = c_1, \sigma(a_1b_1c_1) = e_1$, while $\sigma(a_2d_2)$ and $\sigma(a_1b_1c_2)$ are left undefined by Steps 1 and 2.

STEP 3. Complete $\sigma$ in an arbitrary way.[48]

Because of Step 1, $\zeta(\alpha) = f_\sigma(h)$, for every $h \preceq \zeta(\alpha)$. We want to show that $\sigma$ is a Nash equilibrium. Suppose not. Then there is a decision history $h$ with $h \prec \zeta(\alpha)$ such that, by changing her choice at $h$ from $\sigma(h)$ to a different choice, player $\iota(h)$ can increase her payoff (recall the assumption that the game satisfies the no-consecutive-moves assumption and thus there are no successors of $h$ that belong to player $\iota(h)$). Let $\sigma(h) = a$ (thus $ha \preceq \zeta(\alpha)$) and let $b$ be the choice at $h$ that yields a higher payoff to player $\iota(h)$; that is,

$$u_{\iota(h)}(f_\sigma(hb)) > u_{\iota(h)}(\zeta(\alpha)). \tag{2}$$

Let $\omega \in \mathcal{B}_h(\alpha)$ be such that $hb \preceq \zeta(\omega)$ (such an $\omega$ exists by Point 4 of Definition 2.2). Since $\alpha \in \mathbf{C}$, for every $\omega' \in \mathcal{B}_h(\alpha)$ such that $hb \preceq \zeta(\omega')$, $\zeta(\omega) = \zeta(\omega')$. By Step 2 above,

$$\zeta(\omega) = f_\sigma(hb). \tag{3}$$

It follows from (3) that, at state $\alpha$ and history $h$, player $\iota(h)$ believes that if she plays $b$ her payoff will be $u_{\iota(h)}(f_\sigma(hb))$. Since $\alpha \in \mathbf{T}$, $\alpha \in \mathcal{B}_h(\alpha)$, and since $\alpha \in \mathbf{C}$, for every $\omega' \in \mathcal{B}_h(\alpha)$ such that $ha \preceq \zeta(\omega')$, $\zeta(\omega') = \zeta(\alpha)$. Thus, at state $\alpha$ and history $h$, player $\iota(h)$ believes that if she plays $a$ her payoff will be $u_{\iota(h)}(\zeta(\alpha))$. It follows from this and (2) that at $\alpha$ and $h$ player $\iota(h)$ believes that action $b$ is better than action $a$, which implies that $\alpha \notin \mathbf{R}_h$, contradicting the assumption that $\alpha \in \mathbf{R} \subseteq \mathbf{R}_h$. □

Before proving Proposition 2 we need to define the length of a game.

[47] In the model of Figure 9 we have that $\mathbf{R}_\emptyset = \{\delta, \epsilon, \eta, \theta, \lambda\}, \mathbf{R}_{a_1} = \{\delta\}, \mathbf{R}_{a_2} = \{\alpha, \beta\}, \mathbf{R}_{a_1b_1} = \{\eta, \theta\}, \mathbf{R}_{a_1b_1c_1} = \{\epsilon\}, \mathbf{R}_{a_1b_1c_2} = \{\eta\}, \mathbf{R}_{a_2d_2} = \{\alpha\}$ so that $\mathbf{R} = \{\delta\}$. Furthermore, $\mathbf{T} = \{\gamma, \delta\}$ and $\mathbf{C} = \{\delta, \epsilon, \eta, \theta, \lambda\}$. Hence $\mathbf{T} \cap \mathbf{R} \cap \mathbf{C} = \{\delta\}$.

[48] For instance, in the example of Figure 9, one can complete the above-mentioned partial strategy profile by adding $\sigma(a_2d_2) = g_2$ and $\sigma(a_1b_1c_2) = f_1$ (even though $g_2$ and $f_1$ are "irrational" choices).

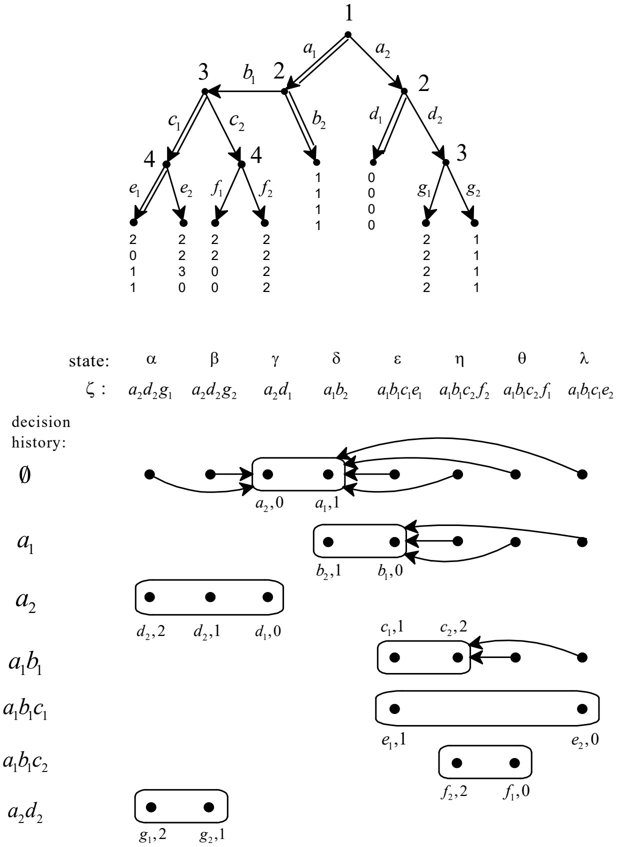

Figure 9: A game and a model of it

**Definition A.2.** The *length of a history* $h$, denoted by $L(h)$, is defined recursively as follows: $L(\emptyset) = 0$ and, for every $a \in A(h)$, $L(ha) = L(h) + 1$; thus the length of history $h$ is the number of actions in $h$. The *length of a game*, denoted by $\ell$, is the length of a longest history in the game: $\ell = \max_{h \in H}\{L(h)\}$.

### Proof of Proposition 2

*Proof.* **(A)** Fix a perfect-information game $G$ and let the pure-strategy profile $\sigma$ be a backward-induction solution of $G$ (that is, a possible output of the backward-induction algorithm). Consider the model generated by $\sigma$ (Definition A.1); then - by construction - for every terminal history $z$ and every decision history $h$ such that $h \prec z$,

$$\forall z' \in Z,\ z' \in \mathcal{B}_h(z) \text{ if and only if } z' = f_\sigma(ha) \text{ for some } a \in A(h). \tag{4}$$

It follows from this that[49]

$$\forall h \in D,\ \text{ if } z = f_\sigma(h) \text{ then } z \in \mathbf{T}_h \tag{5}$$

and[50]

$$\forall z \in Z \text{ and } \forall h \in D \text{ such that } h \prec z,\ z \in \mathbf{C}_h. \tag{6}$$

Let $z^*$ be the play generated by $\sigma$, that is, $z^* = f_\sigma(\emptyset)$. Since every backward-induction solution is a Nash equilibrium, it follows from Part A of Proposition 1 that $z^* \in \mathbf{T} \cap \mathbf{R} \cap \mathbf{C}$. Let $\ell$ be the length of the game. If $\ell = 1$ there is nothing further to prove. Assume, therefore, that $\ell \geq 2$. We need to show that $z^* \in \mathbf{I}_{TRC}$, that is, that, for every decision history $h = a_1...a_m$ $(m \geq 1)$ and for every sequence $\langle z_0, z_1, ..., z_m \rangle$ that leads from $z^*$ to $h$ (see Definition 5.1), $z_m \in \mathbf{T}_h \cap \mathbf{R}_h \cap \mathbf{C}_h$; however, by (6), we only need to show that $z_m \in \mathbf{T}_h \cap \mathbf{R}_h$. Let $h = a_1...a_m$ $(m \geq 1)$ be a decision history and let $\langle z_0, z_1, ..., z_m \rangle$ be a sequence that leads from $z^*$ to $h$ (such a sequence exists: see Remark 3). By Definition 5.1, $z_m \in \mathcal{B}_{a_1...a_{m-1}}(z_{m-1})$, so that, by (4), $z_m = f_\sigma(a_1...a_{m-1}b)$ for some $b \in A(a_1...a_{m-1})$; hence $a_1...a_{m-1}b \preceq z_m$. Again by Definition 5.1, $a_1...a_{m-1}a_m = h \preceq z_m$ and thus $b = a_m$ so that

$$z_m = f_\sigma(h). \tag{7}$$

[49] Proof. Let $h \in D$ and $z = f_\sigma(h)$. Let $a = \sigma(h)$ be the action prescribed by $\sigma$ at $h$. Then $f_\sigma(h) = f_\sigma(ha)$. By (4), $f_\sigma(ha) \in \mathcal{B}_h(z)$ and thus $z \in \mathcal{B}_h(z)$, that is, $z \in \mathbf{T}_h$.

[50] Proof. Let $h \in D$ and $z \in Z$ be such that $h \prec z$. Fix an arbitrary $a \in A(h)$ and arbitrary $z', z'' \in \mathcal{B}_h(z)$ be such that $ha \preceq z'$ and $ha \preceq z''$. Then, by (4), $z' = z'' = f_\sigma(ha)$; hence $z \in \mathbf{C}_h$.

Hence, by (5), $z_m \in \mathbf{T}_h$.

Let $a \in A(h)$ be the action taken at $h$ at state $z_m$ (that is, $ha \preceq z_m$). It follows from (7) that $a = \sigma(h)$. Furthermore, by (4), for every $z' \in \mathcal{B}_h(z_m)$, $z' = f_\sigma(ha')$ for some $a' \in A(h)$. Hence at state $z_m$ and history $h$ player $\iota(h)$ believes that after any choice $a'$ at $h$ the outcome will the one generated by $\sigma$ starting from $ha'$ (that is, the backward-induction outcome induced by $\sigma$ in the subtree that starts at history $ha'$). Furthermore, by (7), the action that she takes at $h$ is $\sigma(h)$, the backward-induction choice prescribed by $\sigma$. Hence player $\iota(h)$ is rational at $h$ and $z_m$, that is, $z_m \in \mathbf{R}_h$.

**(B)** Let $G$ be a perfect-information game. Consider a model of $G$ and a state $\alpha$ in that model such that $\alpha \in (\mathbf{T} \cap \mathbf{R} \cap \mathbf{C}) \cap \mathbf{I}_{TRC}$. We want to show that $\zeta(\alpha)$ is a backward-induction outcome. Let $\ell$ be the length of the game. If $\ell = 1$ then every successor of $\emptyset$ (the root of the tree) is a terminal history. Hence, since $\alpha \in \mathbf{R} \subseteq \mathbf{R}_\emptyset$, the action chosen at $\emptyset$ at state $\alpha$ maximizes player $\iota(\emptyset)$'s payoff and thus is a backward-induction choice. Assume, therefore, that $\ell \geq 2$.

STEP 1. First we show that, at every decision history of length $\ell - 1$ that is reachable from $\alpha$, the action chosen there is a backward-induction choice. Fix an arbitrary decision history $h = a_1 ... a_{\ell-1}$ of length $\ell - 1$ (thus every successor of $h$ is a terminal history) and let $\langle \omega_0, \omega_1, ..., \omega_{\ell-1} \rangle$ be a sequence in $\Omega$ that leads from $\alpha$ to $h$ (such a sequence exists: see Remark 3), that is, (1) $\omega_0 = \alpha$, (2) for every $i = 1, ..., \ell - 1$, $a_1 ... a_i \prec \zeta(\omega_i)$, (3) $\omega_1 \in \mathcal{B}_\emptyset(\alpha)$ and, for every $i = 2, ..., \ell - 1$, $\omega_i \in \mathcal{B}_{a_1 ... a_{i-1}}(\omega_{i-1})$. Since, by hypothesis, $\alpha \in \mathbf{I}_{TRC}$, $\omega_{\ell-1} \in \mathbf{R}_h$ and thus if $b$ is the action taken at history $h$ at state $\omega_{\ell-1}$ (that is, $\zeta(\omega_{\ell-1}) = hb$), then $b$ maximizes the payoff of player $\iota(h)$, that is, $b$ is a backward-induction choice at $h$.

STEP 2. Next we show that, at every decision history of length $\ell - 2$, the active player believes that, for every $a \in A(h)$, if $ha$ is a decision history then the action chosen at $ha$ is a backward-induction action. Fix an arbitrary decision history $h = a_1 ... a_{\ell-2}$ of length $\ell - 2$ and let $\langle \omega_0, \omega_1, ..., \omega_{\ell-2} \rangle$ be a sequence in $\Omega$ that leads from $\alpha$ to $h$ (see Remark 3). Let $a \in A(h)$ be such that $ha$ is a decision history and let $\omega \in \mathcal{B}_h(\omega_{\ell-2})$ be such that $ha \preceq \zeta(\omega)$ (such an $\omega$ exists by Point 4 of Definition 2.2). Then the sequence $\langle \omega_0, \omega_1, ..., \omega_{\ell-2}, \omega \rangle$ reaches $ha$ from $\alpha$ and thus, by Step 1, the action chosen by the active player at $ha$ is a backward-induction action (that is, if $\zeta(\omega) = hab$, with $b \in A(ha)$, then $b$ is a backward-induction choice at $ha$). Furthermore, since $\alpha \in \mathbf{I}_{TRC}$, $\omega_{\ell-2} \in \mathbf{C}_h$ and thus, for every other $\omega' \in \mathcal{B}_h(\omega_{\ell-2})$ such that $ha \preceq \zeta(\omega')$, $\zeta(\omega') = \zeta(\omega)$ and thus, at $h$ and $\omega_{\ell-2}$, player $\iota(h)$ believes that if she takes action $a$ at $h$ then the ensuing outcome is backward-induction outcome $\zeta(\omega)$. From $\alpha \in \mathbf{I}_{TRC}$ it also follows that $\omega_{\ell-2} \in \mathbf{R}_h$ and thus the action chosen by player $\iota(h)$ at $h$ at state $\omega_{\ell-2}$ is optimal given her beliefs that after every choice $a$ at $h$ the outcome following $ha$ is a backward-induction outcome. Finally, from $\alpha \in \mathbf{I}_{TRC}$ it follows that $\omega_{\ell-2} \in \mathbf{T}_h$ so that $\omega_{\ell-2} \in \mathcal{B}_h(\omega_{\ell-2})$ and thus player $\iota(h)$ has correct beliefs at $h$ and at state $\omega_{\ell-2}$ about the outcome following the action actually taken at $h$ and at $\omega_{\ell-2}$ (that is, if $\hat{a}$ is such that $h\hat{a} \preceq \zeta(\omega_{\ell-2})$ then player $\iota(h)$ believes that if she takes action $\hat{a}$ then the outcome will be the backward-induction outcome $\zeta(\omega_{\ell-2})$). Thus $\zeta(\omega_{\ell-2})$ is a backward-induction outcome in the subtree that starts at history $h$.

STEP 3. Iterate the argument of Step 2 backwards to conclude that if $a \in A(\emptyset)$ is decision history of length 1 that is reachable from $\alpha$ via a sequence of the form $\langle \alpha, \beta \rangle$, then $\zeta(\beta)$ is a backward-induction outcome in the subtree that starts at history $a$.

STEP 4. Use the fact that $\alpha \in \mathbf{T}_\emptyset \cap \mathbf{C}_\emptyset$ to conclude that at state $\alpha$ and history $\emptyset$ player $\iota(\emptyset)$ has correct and certain beliefs about the outcomes following decision histories in $A(\emptyset)$ and thus, using the fact that $\alpha \in \mathbf{R}_\emptyset$ and the conclusion of Step 3, deduce that the action taken at state $\alpha$ by $\iota(\emptyset)$ is a backward induction action, so that $\zeta(\alpha)$ is a backward-induction outcome. □

# References

S. Artemov. Robust knowledge and rationality. Technical report, CUNY, 2010. URL http://academicworks.cuny.edu/gc_cs_tr/346/.

R. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

R. Aumann. Reply to binmore. *Games and Economic Behavior*, 17:138–146, 1996.

R. Aumann. On the centipede game. *Games and Economic Behavior*, 23:97–105, 1998.

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

C. Bach and E. Tsakas. Pairwise epistemic conditions for Nash equilibrium. *Games and Economic Behavior*, 85:48–59, 2014.

C. W. Bach and C. Heilmann. Agent connectedness and backward induction. *International Game Theory Review*, 13:195–208, 2011.

D. Balkenborg and E. Winter. A necessary and sufficient epistemic condition for playing backward induction. *Journal of Mathematical Economics*, 27:325–345, 1997.

A. Baltag, S. Smets, and J. Zvesper. Keep hoping for rationality: a solution to the backward induction paradox. *Synthese*, 169:301–333, 2009.

P. Barelli. Consistency of beliefs and epistemic conditions for Nash and correlated equilibria. *Games and Economic Behavior*, 67:363–375, 2009.

P. Battigalli and G. Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53:149–225, 1999.

P. Battigalli, A. Di-Tillio, and D. Samet. Strategies and interactive beliefs in dynamic games. In D. Acemoglu, M. Arellano, and E. Dekel, editors, *Advances in Economics and Econometrics. Theory and Applications: Tenth World Congress, Volume 1*, pages 391–422. Cambridge University Press, Cambridge, 2013.

E. Ben-Porath. Nash equilibrium and backwards induction in perfect information games. *Review of Economic Studies*, 64:23–46, 1997.

K. Binmore. A note on backward induction. *Games and Economic Behavior*, 17: 135–137, 1996.

K. Binmore. Rationality and backward induction. *Journal of Economic Methodology*, 4:23–, 1997.

G. Bonanno. A dynamic epistemic characterization of backward induction without counterfactuals. *Games and Economics Behavior*, 78:31–43, 2013.

G. Bonanno. Reasoning about strategies and rational play in dynamic games. In J. van Benthem, S. Ghosh, and R. Verbrugge, editors, *Models of Strategic Reasoning*, pages 34–62. Springer, 2015.

A. Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.

T. Clausing. Doxastic conditions for backward induction. *Theory and Decision*, 54:315–336, 2003.

T. Clausing. Belief revision in games of perfect information. *Economics and Philosophy*, 20:89–115, 2004.

Y. Feinberg. Subjective reasoning - dynamic games. *Games and Economic Behavior*, 52:54–93, 2005.

I. Gilboa. Can free choice be known? In C. Bicchieri, R. Jeffrey, and B. Skyrms, editors, *The logic of strategy*, pages 163–174. Oxford University Press, 1999.

C. Ginet. Can the will be caused? *The Philosophical Review*, 71:49–55, 1962.

A. Goldman. *A theory of human action*. Princeton University Press, 1970.

J. Halpern. Substantive rationality and backward induction. *Games and Economic Behavior*, 37:425–435, 2001.

M. M. Kaminski. Backward induction and subgame perfection. the justification of a “folk algorithm.”. Technical report, University of California, Irvine, 2009. URL http://www.imbs.uci.edu/files/docs/technical/2009/mbs_09-01.pdf.

M. Ledwig. The no probabilities for acts-principle. *Synthese*, 144:171–180, 2005.

I. Levi. *Hard choices*. Cambridge University Press, 1986.

I. Levi. *The covenant of reason: rationality and the commitments of thought*. Cambridge University Press, 1997.

D. Lewis. *Counterfactuals*. Harvard University Press, 1973.

M. Osborne and A. Rubinstein. *A course in game theory*. MIT Press, Cambridge, 1994.

A. Penta. Robust dynamic mechanism design. Technical report, University of Wisconsin, Madison, 2009. URL http://www.econ.wisc.edu/~apenta/DMD.pdf.

A. Perea. Epistemic foundations for backward induction: an overview. In J. van Benthem, D. Gabbay, and B. Löwe, editors, *Interactive logic. Proceedings of the 7th Augustus de Morgan Workshop*, volume 1 of *Texts in Logic and Games*, pages 159–193. Amsterdam University Press, 2007a.

A. Perea. A one-person doxastic characterization of Nash strategies. *Synthese*, 158:251–271, 2007b.

A. Perea. *Epistemic game theory: reasoning and choice*. Cambridge University Press, Cambridge, 2012.

A. Perea. Belief in the opponents' future rationality. *Games and Economic Behavior*, 83:231–254, 2014.

B. Polak. Epistemic conditions for Nash equilibrium, and common knowledge of rationality. *Econometrica*, 67:673–676, 1999.

A. Quesada. From common knowledge of rationality to backward induction. *International Game Theory Review*, 5:127–137, 2003.

D. Samet. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior*, 17:230–251, 1996.

D. Samet. Common belief of rationality in games of perfect information. *Games and Economic Behavior*, 79:192–200, 2013.

G. L. S. Shackle. *Time in economics*. North Holland Publishing Company, Amsterdam, 1958.

W. Spohn. Where Luce and Krantz do really generalize Savage's decision model. *Erkenntnis*, 11:113–134, 1977.

W. Spohn. *Strategic Rationality*, volume 24 of *Forschungsberichte der DFG-Forschergruppe Logik in der Philosophie*. Konstanz University, 1999. URL http://books.google.com/books?id=4HsLPwAACAAJ.

R. Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in logical theory*, pages 98–112. Blackwell, 1968.

R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

R. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.