forthcoming in Games and Economic Behavior, 78, 2013, 31-45

DOI: http://dx.doi.org/10.1016/j.geb.2012.12.004

A dynamic epistemic characterization of backward induction without counterfactuals ☆

Giacomo Bonanno

Department of Economics
University of California
Davis, CA 95616-8578 - USA

Abstract

We propose a dynamic framework where the rationality of a player's choice is judged on the basis of the actual beliefs that he has at the time he makes that choice. The set of "possible worlds" is given by state-instant pairs (ω, t), where each state specifies the entire play of the game. At every (ω, t) the beliefs of the active player provide an answer to the question "what will happen if I take action a?", for every available action a. A player is rational at (ω, t) if either he is not active or the action he takes is optimal given his beliefs. We characterize backward induction in terms of the following event: the first mover (i) is rational and has correct beliefs, (ii) believes that the active player at date 1 is rational and has correct beliefs, (iii) believes that the active player at date 1 believes that the active player at date 2 is rational and has correct beliefs, etc.

Keywords: perfect-information game, backward induction, dynamic interactive beliefs, rationality, Kripke frame

1. Introduction

The analysis of rational play in dynamic games is usually done within a *static* framework that specifies, for every player, his initial beliefs as well as his disposition to revise those beliefs conditional on hypothetical states of information that the player might find himself in. This is done by means of interactive structures which model a rather complex web of beliefs: for example, Player 2 might initially believe that Player 1 will end the game right away and yet have very detailed beliefs about what Player 1 would believe about Player 2's revised beliefs if Player 1 were instead to give the move to Player 2. In these models each player is assumed to have not only a disposition to revise his own beliefs, should he be faced with unexpected information, but also to have (conditional) beliefs about the disposition of the other players to revise their beliefs. This seems to constitute a rather "heavy" approach to modeling the players' states of mind in

☆I am grateful to two anonymous referees and the Advisory Editor for helpful comments and suggestions. A first draft of this paper was presented at the *First CSLI Workshop on Logic, Rationality and Interaction*, Stanford University, June 2012 and at the *Tenth Conference on Logic and the Foundations of Game and Decision Theory* (LOFT10), University of Sevilla, June 2012.

Email address: gfbonanno@ucdavis.edu (Giacomo Bonanno)
URL: http://www.econ.ucdavis.edu/faculty/bonanno/ (Giacomo Bonanno)

a dynamic game. It is shown in this literature [6, 7, 8, 25, 30] that common *initial* belief of rationality does not imply a backward induction outcome in perfect-information games.

In this paper we suggest an alternative and "lighter" approach, where the rationality of a player's choice is judged on the basis of the *actual beliefs* that the player has *at the time he makes that choice*. We propose a dynamic analysis of perfect-information games where the set of "possible worlds" is given by state-instant pairs (ω, t). Each state ω specifies the entire play of the game and, for every instant t, (ω, t) specifies the history that is reached at that instant (in state ω). A player is said to be *active* at (ω, t) if the history reached in state ω at time t is a decision history of his. At every state-instant pair (ω, t) the beliefs of the active player provide an answer to the question "what will/might happen if I take action a?", for every available action a. A player is said to be rational at (ω, t) if either he is not active there or the action he ends up taking at state ω is optimal given his beliefs at (ω, t). We provide a characterization of backward induction in terms of the following event: the first mover (i) is rational and has correct beliefs, (ii) believes that the active player at date 1 is rational and has correct beliefs, (iii) believes that the active player at date 1 believes that the active player at date 2 is rational and has correct beliefs, etc.

This can be stated more precisely as follows. First we define a time-t belief operator B_t which captures the beliefs of the active player and enables us to express a player's belief that the next player will respond rationally to his choice. Let $\mathbf{T}_t$ be the set of states where the active player at date t (if there is any) has correct beliefs and let $\mathbf{R}_t$ be the set of states where the choice of the active player at date t is rational. In keeping with the literature, we focus on perfect-information games with no relevant ties where there is a unique backward-induction solution. We prove the following characterization. For every m greater than or equal to the depth of the game, if $\omega \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0 (\mathbf{T}_1 \cap \mathbf{R}_1) \cap B_0 B_1 (\mathbf{T}_2 \cap \mathbf{R}_2) \cap ... \cap B_0 B_1 ... B_{m-2} (\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1})$ then the play associated with ω is the backward-induction play. Conversely, if z is the backward-induction play then there is a model of the game and a state ω such that $\omega \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0 (\mathbf{T}_1 \cap \mathbf{R}_1) \cap ... \cap B_0 B_1 ... B_{m-2} (\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1})$ and the play associated with ω is z.

Thus we provide an epistemic characterization of backward induction that does not rely on (objective or subjective) counterfactuals or on dispositional belief revision. Furthermore, strategies do not play any role in our framework.

The analysis is developed in Sections 2 and 3, while Section 4 is devoted to a discussion of conceptual aspects of the proposed approach and of related literature. The proofs are given in the Appendix.

2. Perfect-information games and models

We use the history-based definition of extensive-form game. 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 $1 \leq j \leq k$, the sequence $\left\langle a_1, ..., a_j \right\rangle$ is called a *prefix* of 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.

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.
- A finite set of *histories* $H \subseteq A^*$ which is closed under prefixes (that is, if $h \in H$ and $h' \in A^*$ is a prefix of h, then $h' \in H$). 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. $D = H \backslash Z$ 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*.
- 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. For every $i \in N$, let $D_i = \iota^{-1}(i)$ be the set of histories assigned to player 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).

Given a history $h \in H$, we denoted by $\ell(h)$ the length of h, which is defined recursively as follows: $\ell(\emptyset) = 0$ and if $h \in D$ and $a \in A(h)$ then $\ell(ha) = \ell(h) + 1$. Thus $\ell(h)$ is equal to the number of actions that appear in h; for example, if $h = \langle \emptyset, a_1, a_2, a_3 \rangle$ then $\ell(h) = 3$. We denote by $\ell^{\max}$ the length of the maximal histories in H: $\ell^{\max} = \max_{h \in H}\{\ell(h)\}$. Clearly, if $\ell(h) = \ell^{\max}$ then $h \in Z$. Given a history $h \in H$ and an integer t with $0 \leq t \leq \ell^{\max}$, we denote by h_t the prefix of h of length t. For example, if $h = \langle \emptyset, a, b, c, d \rangle$, then $h_0 = \emptyset$, $h_2 = \langle \emptyset, a, b \rangle$, etc.

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

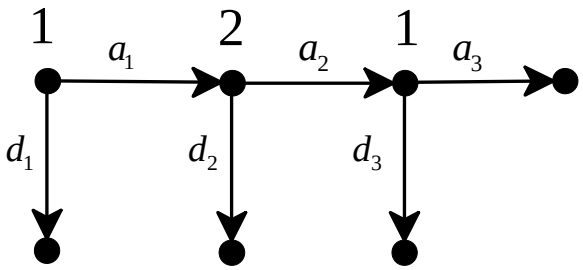
Let Ω be a set of states and $T = \{0, 1, \ldots m\}$ a set of instants or dates. We call $\Omega \times T$ the set of *state-instant pairs*. If $E \subseteq \Omega \times T$ and $t \in T$, we denote by E_t the set of states $\{\omega \in \Omega : (\omega, t) \in E\}$.

Definition 1. Given an extensive form with perfect information $G = \langle A, H, N, \iota \rangle$, a *state-time representation of* G is a triple $\langle \Omega, T, \zeta \rangle$ where Ω is a set of states, $T = \{0, 1, ..., m\}$ with $m \geq \ell^{\max}$ (recall that $\ell^{\max}$ is the depth of the game) and $\zeta : \Omega \to Z$ is a function that assigns to every state a terminal history. Given a state-instant pair $(\omega, t) \in \Omega \times T$, let

$$\zeta_t(\omega) = \begin{cases} \text{the prefix of } \zeta(\omega) \text{ of length } t & \text{if } t < \ell(\zeta(\omega)) \\ \zeta(\omega) & \text{if } t \geq \ell(\zeta(\omega)). \end{cases}$$

Interpretation: the play of the game unfolds over time; the first move is made at date 0, the second move at date 1, etc. A state $\omega \in \Omega$ specifies a particular play of the game (that is, a complete sequence of moves leading to terminal history $\zeta(\omega)$); $\zeta_t(\omega)$ denotes the "state of play at time t" in state ω, that is, the partial history of the play up to date t [if t is less than the length of $\zeta(\omega)$, otherwise - once the play is completed - the state of the system remains at $\zeta(\omega)$].

Figure 1 shows an extensive form with perfect information and a state-time representation of it. For every $\omega \in \Omega = \{\alpha, \beta, \gamma\}$ and $t \in T = \{0, 1, 2, 3\}$ we have indicated the (partial) history $\zeta_t(\omega)$ (recall that $\emptyset$ denotes the empty history). For example, $\zeta_2(\alpha) = a_1 a_2$, $\zeta_2(\beta) = d_1$, etc.



	ζ :	$a_1a_2a_3$	d_1	a_1d_2
	state:	α	β	γ
time:				
0		$\emptyset$	$\emptyset$	$\emptyset$
1		a_1	d_1	a_1
2		a_1a_2	d_1	a_1d_2
3		$a_1a_2a_3$	d_1	a_1d_2

Figure 1: An extensive form with perfect information and a state-time representation of it.

We want to define the notion of rational behavior in a game and examine its implications. Player i chooses rationally at a decision history of his if the choice he makes there is optimal given the beliefs that he holds *at the time at which he makes that choice.* These beliefs might be different from his initial beliefs about what would happen in the game and thus might be revised beliefs in light of the information he has at the moment. However, his prior beliefs are *not relevant* in assessing the rationality of his choice: what counts is what he believes at the time he makes the decision. The beliefs (prior or revised) of the other players are also not relevant. Thus in order to assess the rationality of the actual behavior of the players all we need to specify, at every state-instant pair (ω, t), are the *actual* beliefs of the *active* player. This can be done within a state-time representation of the game, as follows. Given a state ω and an instant t, there will be a unique player who makes a decision at (ω, t) (unless the play of the game has already

reached a terminal history, in which case there are no decisions to be made). If $\zeta_t(\omega)$ is a decision history, the active player is $\iota(\zeta_t(\omega))$; denote $\zeta_t(\omega)$ by h and $\iota(\zeta_t(\omega))$ by i. Then player i has to choose an action from the set $A(h)$. In order to make this choice he will form some beliefs about what will happen if he chooses action a, for every $a \in A(h)$. These beliefs will be used to assess the rationality of the choice that the player ends up making at state ω. We will describe a player's beliefs about the consequences of taking alternative actions by means of an accessibility relation. Thus we use Kripke frames and represent qualitative, rather than probabilistic, beliefs. In order to simplify the notation, we will assign beliefs also to the non-active players, but in a trivial way by making those players believe everything.

We recall the following facts about Kripke frames. If Ω is a set of states and $\mathcal{B}_i \subseteq \Omega \times \Omega$ a binary relation on Ω (representing the beliefs of individual i), for every $\omega \in \Omega$ we denote by $\mathcal{B}_i(\omega)$ the set of states that are reachable from ω using $\mathcal{B}_i$, that is, $\mathcal{B}_i(\omega) = \{\omega' \in \Omega : \omega \mathcal{B}_i \omega'\}$. $\mathcal{B}_i$ is *serial* if $\mathcal{B}_i(\omega) \neq \varnothing$, for every $\omega \in \Omega$; it is *transitive* if $\omega' \in \mathcal{B}_i(\omega)$ implies $\mathcal{B}_i(\omega') \subseteq \mathcal{B}_i(\omega)$ and it is *euclidean* if $\omega' \in \mathcal{B}_i(\omega)$ implies $\mathcal{B}_i(\omega) \subseteq \mathcal{B}_i(\omega')$. Subsets of Ω are called *events*. If $E \subseteq \Omega$ is an event, we say that at $\omega \in \Omega$ individual i believes E if and only if $\mathcal{B}_i(\omega) \subseteq E$. Thus one can define a *belief operator* $B_i : 2^\Omega \to 2^\Omega$ as follows: $B_i E = \{\omega \in \Omega : \mathcal{B}_i(\omega) \subseteq E\}$. Hence $B_i E$ is the event that individual i believes E. It is well known that seriality of $\mathcal{B}_i$ corresponds to consistency of beliefs (if the individual believes E then it is not the case that he believes not E : $B_i E \subseteq \neg B_i \neg E$, where, for every event F, $\neg F$ denotes the complement of F in Ω), transitivity corresponds to positive introspection (if the individual believes E then he believes that he believes E : $B_i E \subseteq B_i B_i E$) and euclideanness corresponds to negative introspection (if the individual does not believe E then he believes that he does not believe E : $\neg B_i E \subseteq B_i \neg B_i E$). [1]

Definition 2. Given an extensive form with perfect information G, a *model of* G is a tuple $\left\langle \Omega, T, \zeta, \{\mathcal{B}_{i,t}\}_{i \in N, t \in T} \right\rangle$ where $\langle \Omega, T, \zeta \rangle$ is a state-time representation of G (see Definition 1) and, for every player $i \in N$ and instant $t \in T$, $\mathcal{B}_{i,t} \subseteq \Omega \times \Omega$ is a binary relation on the set of states (representing the beliefs of player i at time t) that satisfies the following properties: $\forall \omega \in \Omega$,

1. If $i \neq \iota(\zeta_t(\omega))$, that is, if $\zeta_t(\omega)$ is *not* a decision history of player i, then $\mathcal{B}_{i,t}(\omega) = \varnothing$.
2. If $i = \iota(\zeta_t(\omega))$, that is, if $\zeta_t(\omega)$ *is* a decision history of player i, then
   - 2.1. $\mathcal{B}_{i,t}$ is locally serial, transitive and euclidean (that is, $\mathcal{B}_{i,t}(\omega) \neq \varnothing$ and if $\omega' \in \mathcal{B}_{i,t}(\omega)$ then $\mathcal{B}_{i,t}(\omega') = \mathcal{B}_{i,t}(\omega)$).
   - 2.2. If $\omega' \in \mathcal{B}_{i,t}(\omega)$ then $\zeta_t(\omega') = \zeta_t(\omega)$.
   - 2.3. For every $a \in A(\zeta_t(\omega))$ there exists an $\omega' \in \mathcal{B}_{i,t}(\omega)$ such that $\zeta_{t+1}(\omega') = \zeta_t(\omega')a$.

Condition 1 says that a player has trivial beliefs (that is, he believes everything) at all the state-instant pairs where he is not active. We impose this condition only for notational convenience, to eliminate the need to keep track, at every state-instant pair, of who the active player is.[2] To understand Condition 2, fix a state-instant pair (ω, t), let $h = \zeta_t(\omega)$ and suppose that h is a decision history of player i (thus $i = \iota(\zeta_t(\omega))$) where he has to choose an action from the set $A(h)$.

[1] For more details see [5]. We have restricted attention to qualitative, or non-probabilistic, beliefs (represented by binary relations) since they are sufficient for obtaining an epistemic characterization of backward-induction in perfect-information games. In a probabilistic setting the interpretation of the event $B_i E$ would be "the set of states where player i attaches probability 1 to event E".

[2] As explained below, by defining $\mathcal{B}_t = \bigcup_{i \in N} \mathcal{B}_{i,t}$, we can take the relation $\mathcal{B}_t$ to be a description of the beliefs of the active player at date t (whose identity can change from state to state).

Condition 2.1 says that player i has beliefs with standard properties; note that these properties (consistency, positive and negative introspection) are only assumed to hold locally, that is, at state ω.[3] Condition 2.2 says that every state ω' which is accessible from ω by $\mathcal{B}_{i,t}$ (that is, every state that player i considers possible at state ω and instant t) is such that the history associated with (ω', t) is still h; in other words, player i at time t knows that his decision history h has been reached. Condition 2.3 says that for *every* action a available at h, there is a state ω' that player i considers possible ($\omega' \in \mathcal{B}_{i,t}(\omega)$) where he takes action a; that is, the truncation of $\zeta(\omega')$ at time $t+1$ (namely $\zeta_{t+1}(\omega')$) is equal to ha (recall that, by Condition 2.2, $\zeta_t(\omega') = h$). This means that, for every available action, player i has a belief about what will (or might) happen if he chooses that action.

Remark 1. It is worth noting that this way of modeling beliefs is a departure from the standard approach in the literature, where it is assumed that if a player takes a particular action at a state then he knows that he 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. In our approach a player's beliefs refer to the *deliberation* or *pre-choice stage*, where the player considers the consequences of all his actions, without prejudging his subsequent decision.[4] 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 in two different states, say α and β, and be labeled as rational at state α and irrational at state β, because the action he ends up taking at state α is optimal given those beliefs, while the action he ends up taking at state β is not optimal given those same beliefs.

Figure 2 shows a perfect information game and Figure 3 a model of it. We represent a belief relation $\mathcal{B}$ as follows: for any two states ω and ω', $\omega' \in \mathcal{B}(\omega)$ if and only if either ω and ω' are enclosed in the same rounded rectangle or there is an arrow from ω to the rounded rectangle containing ω'.[5] The relations shown in Figure 3 are those of the active players: the relation at date 0 is that of Player 1 ($\mathcal{B}_{1,0}$), the relation at date 1 for states α, β and γ is that of Player 2 ($\mathcal{B}_{2,1}$), the relation at date 1 for states δ, ε and η is that of Player 3 ($\mathcal{B}_{3,1}$) and the relation at date 2 for states α and β is that of Player 3 ($\mathcal{B}_{3,2}$).[6] Consider a state, say α. Then α describes the following beliefs: at date 0 Player 1 believes that if she takes action a_1 then Player 2 will follow (at date 1) with b_2 (state γ) and if she takes action b_1 then Player 3 will follow (at date 1) with either c_2 (state δ) or d_2 (state ε); at date 1 Player 2 (knows that Player 1 played a_1 and) believes that if he takes action a_2 then Player 3 will follow (at date 2) with b_3 (and if he takes action b_2 the game will end). At state α Player 1 ends up playing a_1, Player 2 ends up playing a_2 and Player 3 ends up playing a_3.

[3] Note that local transitivity and euclideanness (positive and negative introspection) are not needed for our results. We have imposed these properties because they are considered in the literature to be necessary properties of "rational" beliefs and because they simplify the graphical representation of beliefs.

[4] Further discussion of this point can be found in Section 4.

[5] In other words, for any two states ω and ω' 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 ω to a rounded rectangle then, for every ω' in the rectangle, $(\omega,\omega') \in \mathcal{B}$.

[6] Thus $\mathcal{B}_{1,0}(\omega) = \{\gamma,\delta,\varepsilon\}$ for every $\omega \in \Omega$, $\mathcal{B}_{2,1}(\omega) = \{\beta,\gamma\}$ for every $\omega \in \{\alpha,\beta,\gamma\}$, $\mathcal{B}_{3,1}(\omega) = \{\delta,\varepsilon,\eta\}$ for every $\omega \in \{\delta,\varepsilon,\eta\}$ and $\mathcal{B}_{3,2}(\omega) = \{\alpha,\beta\}$ for every $\omega \in \{\alpha,\beta\}$. For any remaining state ω and date t, $\mathcal{B}_{i,t}(\omega) = \varnothing$, for every player i (thus, for example, $\mathcal{B}_{1,1}(\omega) = \mathcal{B}_{1,2}(\omega) = \mathcal{B}_{1,3}(\omega) = \varnothing$, for every state ω).

It is worth noting that the notion of model that we are using allows for erroneous beliefs (since we have not imposed reflexivity of the belief relations). Indeed, in the model of Figure 3, at state α Player 1 has incorrect beliefs about the subsequent move of Player 2 if she herself plays a_1 (she believes that Player 2 will follow with b_2 while, in fact, he will play a_2).

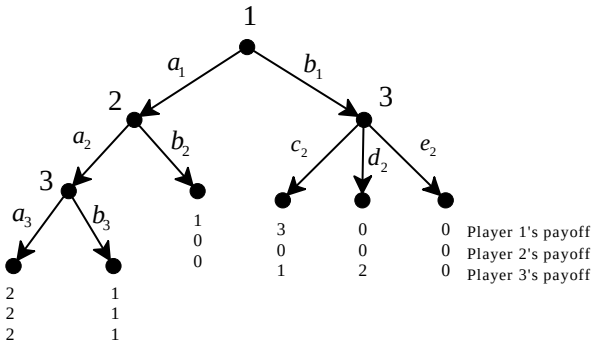


Figure 2: A perfect-information game.

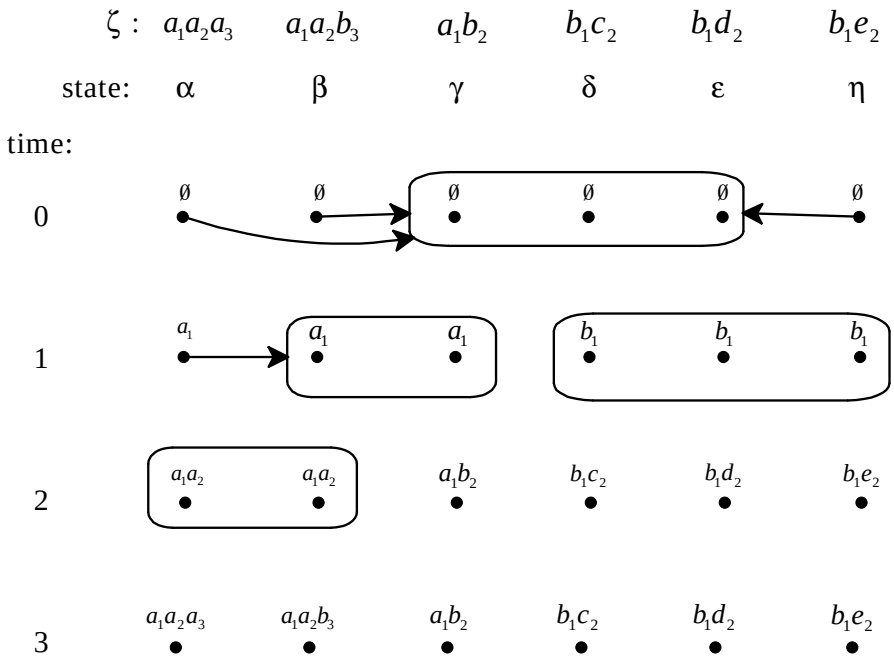


Figure 3: A model of the game of Figure 2.

3. Rationality and backward induction

We shall use a very weak notion of rationality, which has been referred to in the literature as "material rationality" (see, for example, [1, 2, 6, 25]). We say that at a state-instant pair (ω, t) a player is rational if either she is not active at $\zeta_t(\omega)$ (that is, $\zeta_t(\omega)$ is not a decision history of hers) or the action that she ends up choosing at ω is "optimal" given her beliefs at date t, in the sense that it is not the case that - according to her beliefs - there is another action of hers that guarantees higher utility. Thus a player is *irrational* at a state-instant pair (ω, t) if she is active at history $\zeta_t(\omega)$, she ends up taking action a at ω and she believes that her maximum utility if she takes action a is less than the minimum utility that she gets if she takes some other action a'.

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.

Definition 3. Fix an arbitrary player i and an arbitrary state-instant pair (ω, t). We say that player i is *rational at* (ω, t) if either

(1) $\zeta_t(\omega)$ is not a decision history of player i, or

(2) $\zeta_t(\omega)$ is a decision history of player i and if a is the action chosen by player i at ω (that is, $\zeta_{t+1}(\omega) = \zeta_t(\omega)a$) then, for every $a' \in A(\zeta_t(\omega))$, it is not the case that $\min_{\omega' \in A'}\{U_i(\zeta(\omega'))\} > \max_{\omega' \in A}\{U_i(\zeta(\omega'))\}$ where $A' = \{\omega' \in \mathcal{B}_{i,t}(\omega) : \zeta_{t+1}(\omega') = \zeta_t(\omega')a'\}$ and $A = \{\omega' \in \mathcal{B}_{i,t}(\omega) : \zeta_{t+1}(\omega') = \zeta_t(\omega')a\}$ (recall that $U_i : Z \to \mathbb{R}$ is player i's utility function on the set of terminal histories).

For example, in the model of Figure 3, Player 1 is rational at state α and date 0, because she believes that if she takes action a_1 then her payoff will be 1 (she believes that Player 2 will follow with b_2) and if she takes action b_1 then her payoff will be either 3 or 0 (she believes that Player 3 will follow with either c_2 or d_2) and she actually ends up taking action a_1. Similarly, Player 2 is rational at state α and date 1 and Player 3 is rational at state α and date 2. On the other hand, Player 2 is not rational at state γ and date 1 (he believes that if he takes action a_2 his payoff will be 1 and if he takes action b_2 his payoff will be 0 and yet he ends up taking action b_2). Thus, since $\gamma \in \mathcal{B}_{1,0}(\alpha)$, at state α and date 0 it is not the case that Player 1 believes that Player 2 will choose rationally at date 1.

We denote by $\mathbf{R}_t \subseteq \Omega$ the event that (that is, the set of states at which) the active player (if there is one) is rational at date t.[7] Thus $\omega \in \mathbf{R}_t$ if and only if either $\zeta_t(\omega)$ is a terminal history (that is, $\zeta_t(\omega) = \zeta(\omega)$) or $\zeta_t(\omega)$ is a decision history and the active player at $\zeta_t(\omega)$ is rational at (ω, t). Of course, the identity of the active player can vary across states, that is, the active player at (ω, t) can be different from the active player at (ω', t). In the model of Figure 3 we have that $\mathbf{R}_0 = \Omega$, $\mathbf{R}_1 = \{\alpha, \beta, \varepsilon\}$, $\mathbf{R}_2 = \{\alpha, \gamma, \delta, \varepsilon, \eta\}$ and $\mathbf{R}_3 = \Omega$.

Let $B_{i,t} : 2^\Omega \to 2^\Omega$ be the belief operator of player i at date t. Thus, for every event $E \subseteq \Omega$, $B_{i,t}E = \{\omega \in \Omega : \mathcal{B}_{i,t}(\omega) \subseteq E\}$. By Condition 1 of Definition 2, if player i is not active at (ω, t) then $\mathcal{B}_{i,t}(\omega) = \varnothing$ and thus $\omega \in B_{i,t}E$ for every event E. Let $B_t : 2^\Omega \to 2^\Omega$ be the operator defined by $B_tE = \bigcap_{i \in N} B_{i,t}E$ (thus $\omega \in B_tE$ if and only if $\bigcup_{i \in N} \mathcal{B}_{i,t}(\omega) \subseteq E$). Then B_tE is the event that "the active player believes E at time t" (which is trivially equivalent to the event that "everybody believes E at time t"). We summarize this in the following remark.

[7] By Definition 3 inactive players are always rational; thus $\mathbf{R}_t$ can also be described as the event that "every player is rational at date t".

Remark 2. For every $\omega \in \Omega$ and $t \in T$, define $\mathcal{B}_t(\omega) = \bigcup_{i \in N} \mathcal{B}_{i,t}(\omega)$ and $B_t : 2^\Omega \to 2^\Omega$ by $B_t E = \bigcap_{i \in N} B_{i,t} E$ (thus $\omega \in B_t(E)$ if and only if $\mathcal{B}_t(\omega) \subseteq E$.) It follows that if j is the active player at $\zeta_t(\omega)$, then $\mathcal{B}_t(\omega) = \mathcal{B}_{j,t}(\omega)$ and, for every event E, $\omega \in B_t(E)$ if and only if $\mathcal{B}_{j,t}(\omega) \subseteq E$.

For example, in the model of Figure 3, we have that $\alpha \notin B_0 \mathbf{R}_1$ (since $\gamma \in \mathcal{B}_0(\alpha)$ and $\gamma \notin \mathbf{R}_1$), that is, it is not the case that the active player at date 0 (Player 1) believes that the active player at date 1 will choose rationally. Indeed, at date 0 Player 1 believes that if she plays a_1 then the active player at date 1 (Player 2) will not choose rationally and if she plays b_1 then the active player at date 1 (Player 3) might or might not choose rationally (Player 3 chooses rationally at ε but not at δ).

Note that the models we are considering allow for the possibility that a player may ascribe to a future mover beliefs that are different from the beliefs that that player will actually have. In other words, a player may have erroneous beliefs about the future beliefs of other players (or even about her own future beliefs).

Let $\mathbf{T}_t$ be the set of states where the beliefs of the active player (if there is one) are correct: $\mathbf{T}_t = \{\omega \in \Omega : \textit{if } \mathcal{B}_t(\omega) \neq \varnothing \textit{ then } \omega \in \mathcal{B}_t(\omega)\}$.[8] For example, in the model of Figure 3 we have that $\mathbf{T}_0 = \{\gamma, \delta, \varepsilon\}$, $\mathbf{T}_1 = \{\beta, \gamma, \delta, \varepsilon, \eta\}$ and $\mathbf{T}_2 = \mathbf{T}_3 = \Omega$. Thus if $\omega \in \mathbf{T}_t$ and $\zeta_t(\omega)$ is a decision history, then, for every event $E \subseteq \Omega$, if the active player believes E (that is, if $\mathcal{B}_t(\omega) \subseteq E$) then E is indeed true (that is, $\omega \in E$).

In keeping with the literature, we restrict attention to games without relevant ties.

Definition 4. 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 $U_i(z) \neq U_i(z')$.

For example, the game shown in Figure 2 has no relevant ties. If a game has no relevant ties, then it has a unique backward-induction solution.[9]

The following two propositions (which are proved in the Appendix) provide a characterization of backward induction in terms of the following event (where $m = \ell^{max}$ is the depth of the game, that is, the length of its maximal histories):

$$(\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0 (\mathbf{T}_1 \cap \mathbf{R}_1) \cap B_0 B_1 (\mathbf{T}_2 \cap \mathbf{R}_2) \cap ... \cap B_0 B_1 ... B_{m-2} (\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1}) \qquad (1)$$

The interpretation of (1) is: the set of states where the active player at time zero (i) is rational and has correct beliefs and (ii) believes that the active player at time 1 is rational and has correct beliefs and (iii) believes that the active player at time 1 believes that the active player at time 2 is rational and has correct beliefs, and so on. [10]

[8] Thus an active player has correct beliefs whenever the true (or actual) state is among those that he considers possible. On the other hand, by definition, inactive players always have correct beliefs. The expression "correct beliefs" is common in the literature. An alternative expression is "non-deluded beliefs". In a probabilistic setting a player has correct beliefs at state ω if he attaches positive probability to ω.

[9] The definition of backward-induction solution is reviewed in the Appendix.

[10] In a probabilistic setting (1) would be interpreted as the event that the active player at time zero (i) is rational and has correct beliefs and (ii) assigns probability 1 to the event that the active player at time 1 is rational and has correct beliefs and (iii) assigns probability 1 to the event that the active player at time 1 assigns probability 1 to the event that the active player at time 2 is rational and has correct beliefs, etc.

Proposition 1. *Fix a perfect-information game G without relevant ties and let m be its depth. Fix an arbitrary model of G and an arbitrary state ω. If $\omega \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0(\mathbf{T}_1 \cap \mathbf{R}_1) \cap B_0B_1(\mathbf{T}_2 \cap \mathbf{R}_2) \cap ... \cap B_0B_1...B_{m-2}(\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1})$ then $\zeta(\omega)$ is the backward-induction terminal history.*

Proposition 2. *Fix a perfect-information game G without relevant ties and let m be its depth. Let z be the backward-induction terminal history. Then there is a model of G and a state ω such that (1) $\zeta(\omega) = z$ and (2) $\omega \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0(\mathbf{T}_1 \cap \mathbf{R}_1) \cap ... \cap B_0B_1...B_{m-2}(\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1})$.*

Remark 3. An apparent shortcoming of the above characterization of backward-induction is that it is stated in terms of the depth of the game and is thus based on a condition that differs across games. However, implicit in the above characterization is a criterion that is uniform across all games. To see this, note that - in any model of a game - for every state ω and for every date t with $t \geq \ell^{\max}$ (recall that $\ell^{\max}$ is the depth of the game), it is trivially the case that $\mathbf{T}_t = \mathbf{R}_t = \Omega$ and thus $B_0B_1...B_{t-1}(\mathbf{T}_t \cap \mathbf{R}_t) = \Omega$. It follows that the characterization given in the above propositions is equivalent to a characterization in terms of the following event S^∞, which provides a uniform condition across all games:

$$S^\infty = (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0(\mathbf{T}_1 \cap \mathbf{R}_1) \cap B_0B_1(\mathbf{T}_2 \cap \mathbf{R}_2) \cap \ ... \ \cap \ B_0B_1...B_{t-1}(\mathbf{T}_t \cap \mathbf{R}_t) \cap \ ...$$

Thus the characterization of backward induction given in Propositions 1 and 2 can be restated as follows, considering now models in which $T = \mathbb{N}$ (where $\mathbb{N}$ denotes the set of natural numbers):
(1) Let G be a perfect-information game without relevant ties and fix an arbitrary model of it and an arbitrary state ω. If $\omega \in S^\infty$ then $\zeta(\omega)$ is the backward-induction terminal history.
(2) Given a perfect-information game G without relevant ties, let z be the backward-induction terminal history. Then there is a model of G and a state ω such that $\zeta(\omega) = z$ and $\omega \in S^\infty$.

The condition in Proposition 1 that beliefs be locally correct is essential. For example, if $\omega \notin \mathbf{T}_0$ then it may happen that $\zeta(\omega)$ is not the backward-induction terminal history even if the other conditions hold. This is shown in Figure 4, where $\mathbf{R}_0 = \{\alpha, \beta\}$, $\mathbf{R}_1 = \{\beta, \gamma\}$, $\mathbf{T}_0 = \{\beta, \gamma\}$, $\mathbf{T}_1 = \Omega$, $B_0\mathbf{R}_1 = B_0\mathbf{T}_1 = \Omega$. Hence $\alpha \in \mathbf{R}_0 \cap B_0(\mathbf{T}_1 \cap \mathbf{R}_1)$ (in this game $\ell^{\max} = 2$) and yet $\zeta(\alpha) = a_1a_2$ which is not the backward-induction play. At state α Player 1 is rational, believes that after her move Player 2 will be rational and will have correct beliefs and yet the play associated with α is not the backward-induction play (because Player 1 is wrong in her belief that Player 2 will play rationally at date 1). Similar examples can be constructed to show that in Proposition 1 the condition $\omega \in B_0\mathbf{T}_1$ is necessary and so is $\omega \in B_0B_1\mathbf{T}_2$, etc.

One may wonder if the following is the case. Consider a state ω where the active player at time 0 chooses rationally, believes at time 0 that the active player at time 1 chooses rationally, believes at time 0 that the active player at time 1 believes that the active player at time 2 chooses rationally, and so on (that is, $\omega \in \mathbf{R}_0 \cap B_0\mathbf{R}_1 \cap B_0B_1\mathbf{R}_2 \cap \ldots$); is it true that, at state ω, the active player at time 0 believes that after each of her possible actions the opponents will choose in accordance with backward induction?[11] The answer is negative, as shown in Figure 5. Here we have that $\mathbf{R}_0 = \Omega, \mathbf{R}_1 = \{\gamma, \delta, \epsilon\}, \mathbf{R}_2 = \{\beta, \gamma, \epsilon\}, B_0\mathbf{R}_1 = B_1\mathbf{R}_2 = B_0B_1\mathbf{R}_2 = \Omega$. Hence $\epsilon \in \mathbf{R}_0 \cap B_0\mathbf{R}_1 \cap B_0B_1\mathbf{R}_2$ (in this game $\ell^{\max} = 3$) and yet at state ϵ it is not the case that the active player at time 0 (Player 1) believes that after playing a_1 Players 2 and 3 will follow with the choices prescribed by the backward-induction solution: $\delta \in \mathcal{B}_0(\epsilon)$ and $\zeta(\delta) = a_1a_2a_3$ while the backward-induction choices after a_1 are a_2 and b_3.[12]

[11] This point was raised by a referee.

[12] Note that in this model we also have that $\epsilon \in \mathbf{R}_1 \cap \mathbf{R}_2 \cap \mathbf{T}_0 \cap \mathbf{T}_1 \cap \mathbf{T}_2 \cap B_0B_1\mathbf{T}_2$; thus $\epsilon \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0\mathbf{R}_1 \cap$

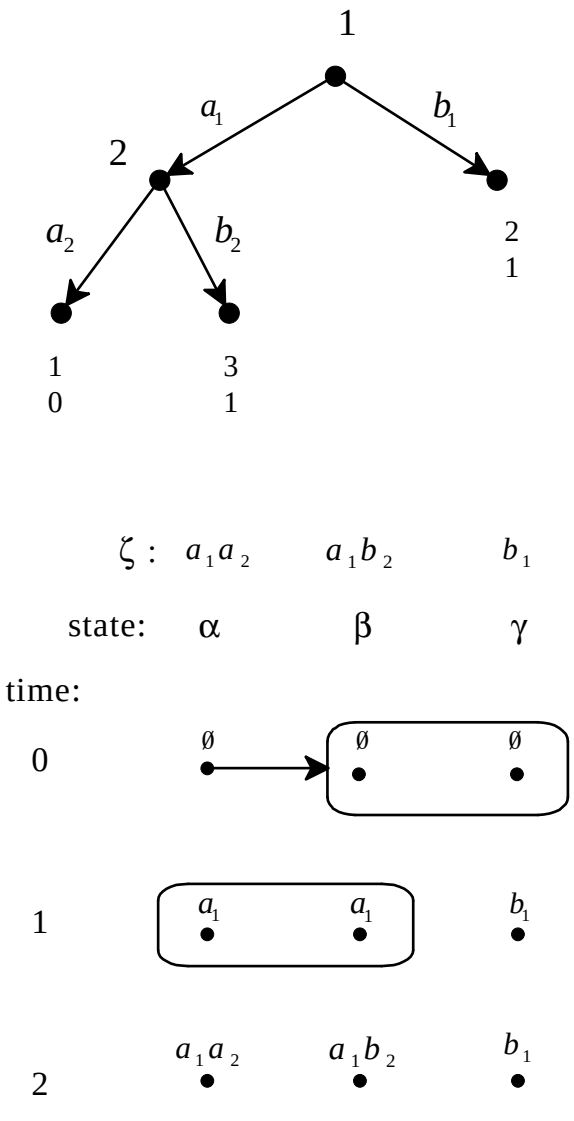


Figure 4: A perfect-information game and a model of it.

$B_0B_1(\mathbf{T}_2 \cap \mathbf{R}_2)$. However, $\epsilon \notin B_0\mathbf{T}_1$.

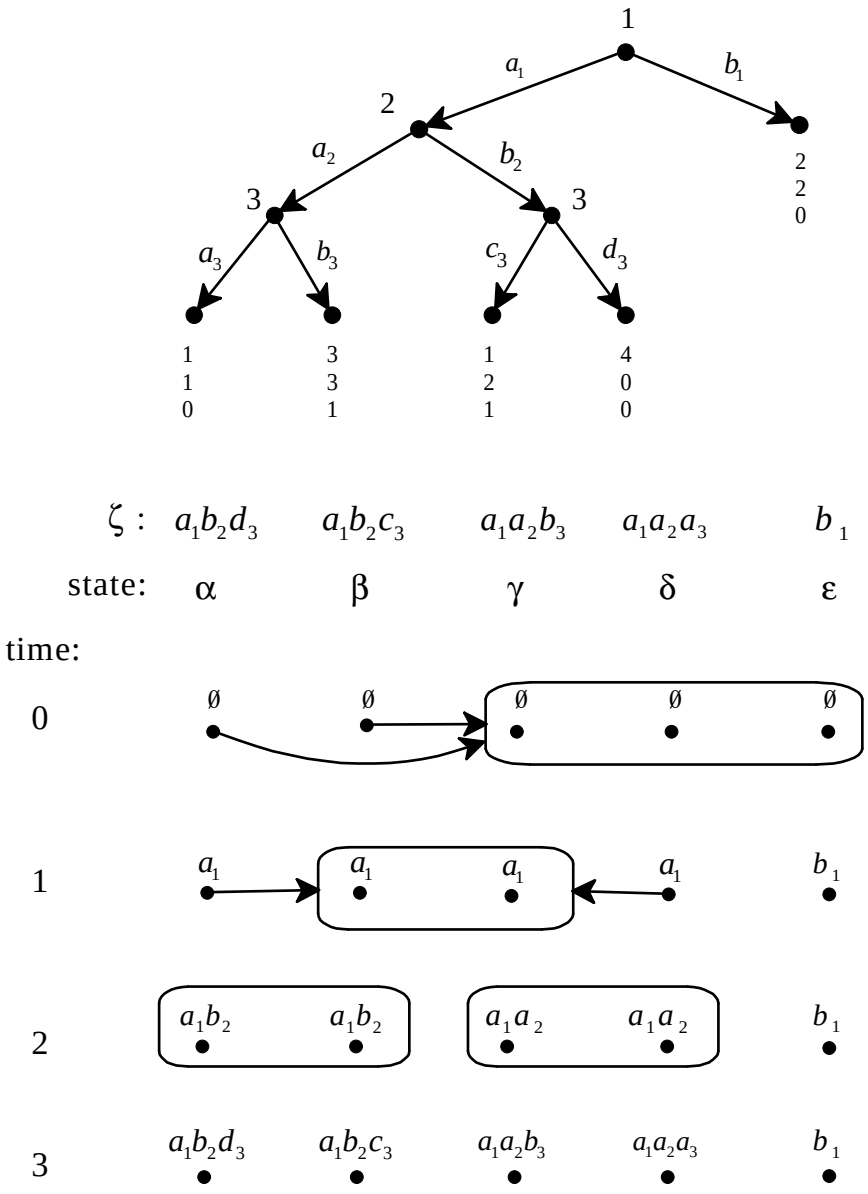


Figure 5: A perfect-information game and a model of it.

4. Discussion and related literature

We have provided a characterization of backward induction in a framework where players' beliefs are modeled as "pre-choice" or "deliberation-stage" beliefs. When it is his turn to move, a player considers the consequences of all his actions, without pre-judging his subsequent decision; in other words, the beliefs of the active player at a state-instant pair are truly open to the possibility of taking any of the available actions. As noted in Remark 1, this constitutes a departure from the standard approach in the literature where it is assumed that if, at a state, a player takes action a, then she knows that she takes action a. As pointed out by several authors, it is the essence of deliberation that one cannot reason towards a choice if one already knows what that choice will be. For instance, Shackle [26, p. 21] remarks that if an agent could predict the option he will choose, his decision problem would be "empty", Ginet [14, p. 50] claims that "it is conceptually impossible for a person to know what a decision of his is going to be before he makes it", Goldman [15, p. 194] writes that "deliberation implies some doubt as to whether the act will be done", Levi states that "the deliberating agent cannot, before choice, predict how he will choose" [18, p. 65] and coins the phrase "deliberation crowds out prediction" [19, p. 81].[13]

While the standard approach in the literature is to model a player's beliefs *after* she has made her choice, we have chosen to model pre-choice beliefs. A potential objection to the proposed approach arises in games where a player chooses more than once along a given play. Consider a model of such a game, a state ω and two instants t_1 and t_2, with $t_1 < t_2$, such that player i moves at both t_1 and t_2 along the play associated with ω. The proposed approach then requires player i at time t_1 to have "open" beliefs about his choice at the decision history associated with (ω, t_1), while allowing him to have beliefs about (or be certain of) what choice he will make at the later time t_2 (that is, at the history associated with (ω, t_2)).[14] This is an issue that has been addressed in the literature and several authors have maintained that there is no inconsistency between the principle that one should not attribute to a player beliefs about his current choice and the claim that, on the other hand, one *can* attribute to the player beliefs about later choices. 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 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." [13, pp. 171-172]

In a similar vein, Levi [19, p. 81] writes that "agent X may coherently assign unconditional credal probabilities to hypotheses as to what he will do when some future opportunity for choice arises. Such probability judgments can have no meaningful role, however, when the opportunity

[13] Similar observations were made by several other authors. For a list of relevant references see [17].

[14] A referee found this possibility a weakness of the proposed approach.

of choice becomes the current one." Similarly, Spohn [27, p. 114] states the principle that "any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts" and then adds [28, pp. 44-45] that in the case of sequential decision making, the decision maker *can* ascribe subjective probabilities to his future (but not to his present) actions. We share the point of view expressed by these authors. If a player moves sequentially at times t_1 and t_2, with $t_1 < t_2$, then at time t_1 he has full control over his immediate choices (those available at t_1) but not over his later choices (those available at t_2). The agent can predict - or form an intention about - his future behavior, but he cannot *irrevocably* decide it, just as he can predict - but not decide - how other players will behave after his current choice.[15]

There is a large literature on the epistemic foundations of backward induction, which was recently reviewed in [9, 21]. In what follows we shall try to highlight the important differences between our approach and the existing literature.

We have focused on a purely behavioral framework where, at each state of the world and at each instant, only the actions actually taken and the beliefs actually held by the active player are specified. Thus, contrary to a well-established literature [1, 2, 3, 7, 8, 16, 20, 22, 23, 24, 29, 30], strategies (or plans of action) do not play any role in our analysis. Indeed the use of strategies in models of dynamic games involves the implicit use of counterfactuals.[16] Methodologically, this is not satisfactory: if it is necessary to specify what a player would do in situations that do not arise in the state under consideration, then one should model the counterfactual explicitly.

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 [25]. Unlike the other papers that take a purely behavioral point of view [4, 6, 7, 25], our analysis does not make use of subjective counterfactuals. The use of subjective counterfactuals or dispositional belief revision is made necessary in that literature by two characteristics of the models used. First of all, the static nature of the framework makes it impossible to model explicitly the beliefs of the players at the time of choice; one thus needs to do so indirectly by representing simultaneously a player's initial beliefs and his disposition to revise those beliefs subject to conceivable items of information that he might receive during the play of the game. This is done either probabilistically using conditional probability systems [6, 7] or by means of qualitative belief revision structures [4, 29, 30]. As pointed out by Stalnaker,

> "It should be noted that even with the addition of the belief revision structure to the epistemic models ..., they remain static models. A model of this kind represents only the agent's beliefs at a fixed time, together with the policies or dispositions to revise her beliefs that she has at that time. The model does not represent any actual revisions that are made when new information is actually received." [31, p. 198].[17]

[15] An implication of this point of view is that, since - at the time of deliberation - the agent does not know what choice he is going to make, he cannot know that his forthcoming choice is rational. For example, as the Advisory Editor pointed out, in the event which characterizes backward induction (given by (1)), at time t the active player doesn't know that she is rational, even though she believes that every future player (and possibly her future self) is rational. This is unavoidable if one wants to model pre-choice or deliberation-stage beliefs. This issue has been discussed at length in the philosophical literature (see, for example, [18, 19]).

[16] While in a simultaneous game the association of a strategy of player i to a state can be interpreted as a description of player i's behavior at that state, in the case of dynamic games this interpretation is no longer valid, since one ends up describing not only the actual behavior of player i but also his counterfactual behavior at decision histories that are not reached in the actual state.

[17] The author goes on to say that "The models can be enriched by adding a temporal dimension to represent the

The second characteristic of the models that use subjective counterfactuals is that they impose the constraint that, if at a state a player takes action a, then she knows that she takes action a (that is, at every state that the player considers possible, she takes action a); thus one needs to use counterfactuals in order to represent a player's beliefs about the consequences of taking an action different from a.

The characterization of backward induction that we have provided is in terms of the forward beliefs of the first mover: she believes in the rationality and correct beliefs of future movers and believes that they, too, will believe in the rationality and correct beliefs of future movers. That this type of condition (belief in future movers' rationality) is central to backward induction is now well understood [3, 4, 10, 11, 12, 23, 30]. The novelty of our approach lies in (i) the switch to a dynamic framework with "pre-choice" beliefs, (ii) showing that the notion of backward induction does not require the use of (objective or subjective) counterfactuals and (iii) pointing out the need for what could be called "local knowledge", that is, locally correct beliefs (as captured by the events $\mathbf{T}_t$).[18]

Since the literature on the epistemic foundations of backward induction has been thoroughly reviewed by Perea [21] it is unnecessary to go into the details of each contribution. We shall therefore close with a few comments on important differences between our characterization of backward induction and those that appear to be conceptually closest, namely [3, 25]. Balkenborg and Winter [3] define a condition that they call "forward knowledge of rationality" at the root of the tree and show that it is sufficient for backward induction. Their condition is conceptually very close to ours, but the framework that they use differs in many crucial respects. First of all, they use Aumann's [1] framework, which is static and partitional (so that players' beliefs are assumed to be *necessarily* correct: see Footnote 18).[19] Secondly, their analysis relies heavily on counterfactuals: they describe states in terms of strategy profiles (so that a state specifies players' behavior also at nodes that are not reached in that state) and use a stronger notion of rationality than ours, namely Aumann's [1] notion of "substantive" rationality, according to which a player can be irrational at a state even if she is not active at that state (that is, even if none of her decision nodes is reached at that state). Finally, it should be noted that the authors restrict attention to games where each player has only *one* decision node.

The behavior-based model (which does not describe states in term of strategies) was introduced by Samet [25]. His analysis relies on subjective counterfactuals, expressed in terms of

---

dynamics, but doing so requires that the knowledge and belief operators be time indexed." In our models the belief operators are indeed time indexed and represent the actual beliefs of the players when actually informed that it is their turn to move.

[18] A strand in the literature [1, 2, 3] assumes that each belief relation is reflexive everywhere, so that it gives rise to a partition of the set of states. In such cases it is common to speak of *knowledge* rather than belief. As Stalnaker points out, it is methodologically preferable to carry out the analysis in terms of (possibly erroneous) beliefs and then - if desired - add further conditions (such as the local correctness of beliefs, that is, local reflexivity of the belief relations). The reason why one should not start with the assumption of *necessarily* correct beliefs (that is, *global* reflexivity of the belief relations) is that such an 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." [29, p. 153].

[19] Quesada [24] shows that the assumptions on beliefs (reflexivity, transitivity and euclideanness) can be relaxed, provided that one strengthens the forward rationality condition by assuming it to hold not only at the root of the tree but also at every other node.

hypothetical knowledge operators.[20] The characterization of backward induction that he provides is in terms of the condition of *common hypothesis of node rationality*, which turns out to have some similarity with our condition.[21] However, Samet's framework differs substantially from ours: his approach is static ("time is absent from the model - we analyze the game at a point in time before it is played" [25, p.233]), it assumes knowledge, rather than belief (see Footnote 18), it imposes the condition that if at a state a player considers it possible that he takes an action at a decision history then he knows that, if that decision history is reached, he takes that action and, finally - as noted above - he makes essential use of subjective counterfactuals, in the form of conditional knowledge operators.

We conclude by noting that the notion of belief in forward rationality has also been extended beyond games with perfect information [20, 22, 23].[22] This is done in a static framework that describes states in terms of strategies (or plans of action).

Appendix A. Proofs

We give below the proofs of Propositions 1 and 2. First we recall the definition of *backward induction solution*. The backward induction solution of a perfect-information game without relevant ties is unique and is given by the output of the following algorithm:

1. Start at a decision history h whose immediate successors are only terminal histories, that is, for every $a \in A(h)$, $ha \in Z$ (e.g. history b_1 in the game of Figure 2) and select the choice that maximizes the utility of player $\iota(h)$ (in the game of Figure 2, at b_1 player 3's utility-maximizing choice is d_2). Delete the successors of h, thus turning h into a terminal history, and assign to h the payoff vector associated with the selected choice.
2. Repeat Step 1 in the reduced game until all the decision histories have been exhausted.

The output of the backward-induction algorithm can be written in terms of a profile of strategies, where a strategy of player i is defined as a list of choices, one for each decision history of player i. For example, the backward induction solution of the game of Figure 2 can be written as $(a_1, a_2, (a_3, d_2))$.

In order to prove Proposition 1 we need the following definition.

Definition 5. Fix a perfect-information game and a model of it. Let $\alpha, \beta \in \Omega$. We say that β *is reachable from* α *with* s *steps* ($s \geq 1$) if there is a sequence of state-instant pairs $\langle(\omega_0, 0), (\omega_1, 1), ..., (\omega_s, s)\rangle$ such that:

[20] A hypothetical knowledge operator $K_i(H, E)$ assigns to each pair of events, H (the hypothesis) and E, the set of states where player i hypothesizes that if H were true, then he would know E.

[21] As the author explains,

> [...] a common hypothesis of node rationality depends only on the hypotheses of the root player. The reason [...] is as follows. Clearly, only the hypotheses of the root player, and nothing else, determine his action at the root. What he hypothesizes about the consequence of his actual action at the root is knowledge since the antecedent is true and, moreover, as the root player knows his action, the consequences are indeed true. But among these consequences are the hypotheses of the next player about what is true if his node is reached. Since the next player's node is indeed reached, his hypotheses are knowledge and so on. [25, p. 244]

[22] Perea [22, 23] characterizes the notion of common belief in future rationality in terms of an iterative procedure of elimination of strategies that are dominated at information sets. In a similar vein, Penta [20] relates the notion of "common certainty of future rationality at every history" to a procedure that he calls "backward rationalizability".

1. $\omega_0 = \alpha$,
2. $\omega_s = \beta$,
3. $\forall k = 1, ..., s$, $\omega_k \in \mathcal{B}_{k-1}(\omega_{k-1})$.

For example, in the model of Figure 3, β is reachable from η with 2 steps with the sequence $\langle(\eta, 0), (\gamma, 1), (\beta, 2)\rangle$.[23]

Remark 4. Let E be an event, α a state and suppose that $\alpha \in B_0B_1...B_{s-1}E$. Then for every $\beta \in \Omega$, if β is reachable from α with s steps then $\beta \in E$. [24]

Proof of Proposition 1. Fix a perfect-information game with no relevant ties, so that there is a unique backward induction (BI) solution. Let $f_{BI} : H \to Z$ be the following function: if h is a decision history then $f_{BI}(h)$ is the terminal history that is reached from h by following the backward-induction choices and if z is a terminal history then $f_{BI}(z) = z$.[25] Recall that if $z \in Z$ and $t \in T$, we denote by z_t the prefix of z of length t. Fix a model of the game and suppose that α is a state such that $\alpha \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0 (\mathbf{T}_1 \cap \mathbf{R}_1) \cap B_0B_1 (\mathbf{T}_2 \cap \mathbf{R}_2) \cap ... \cap B_0B_1...B_{m-2} (\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1})$ (where $m = \ell^{max}$ is the depth of the game). We need to show that $\zeta(\alpha) = f_{BI}(\emptyset)$ (recall that $\emptyset$ denotes the empty history). First we show that,

$$\begin{array}{l} \text{For every } t \text{ with } 1 \leq t \leq m-1 \text{ and for every } \beta \in \Omega, \\ \text{if } \beta \text{ is reachable from } \alpha \text{ with } t \text{ steps then } \zeta(\beta) = f_{BI}(\zeta_t(\beta)). \end{array} \tag{A.1}$$

We prove this by induction.

Base step: $t = m - 1$. Fix an arbitrary β which is reachable from α with $m - 1$ steps. If $\zeta_{m-1}(\beta)$ is a terminal history, then $\zeta_{m-1}(\beta) = \zeta(\beta)$ (see Definition 1) and, by definition of $f_{BI}(\cdot)$, $f_{BI}(\zeta(\beta)) = \zeta(\beta)$. Thus $\zeta(\beta) = f_{BI}(\zeta_{m-1}(\beta))$. Suppose, therefore, that $\zeta_{m-1}(\beta)$ is a decision history. Let i be the active player, that is, the player who moves at $\zeta_{m-1}(\beta)$. Fix an arbitrary $\omega \in \mathcal{B}_{m-1}(\beta)$.[26] Then, by Definition 2, $\zeta_{m-1}(\omega) = \zeta_{m-1}(\beta)$. Since the depth of the game is m, after player i's move at $\zeta_{m-1}(\omega)$ the game ends and thus $\zeta_m(\omega) = \zeta(\omega)$.[27] Since $\alpha \in B_0B_1...B_{m-2}\mathbf{R}_{m-1}$, by Remark 4 $\beta \in \mathbf{R}_{m-1}$, that is, player i is rational at state β and time $m - 1$. Hence, the choice made by player i at state β and time $m - 1$ is the payoff-maximizing choice there, that is, $\zeta(\beta) = f_{BI}(\zeta_{m-1}(\beta))$.

Induction step: suppose that (A.1) is true for $t = k$ with $1 < k \leq m - 1$. We want to show that it is true for $t = k - 1$. Fix an arbitrary β which is reachable from α with $k - 1$ steps.

First we show that,

$$\forall \omega \in \mathcal{B}_{k-1}(\beta), \;\; \zeta(\omega) = f_{BI}(\zeta_k(\omega)). \tag{A.2}$$

[23]Note that, if β is reachable from α with s steps, then $\zeta_{s-1}(\beta)$ is a decision history. In fact, we have that $\beta = \omega_s \in \mathcal{B}_{s-1}(\omega_{s-1})$ and thus $\mathcal{B}_{s-1}(\omega_{s-1}) \neq \varnothing$, so that $\zeta_{s-1}(\omega_{s-1})$ is a decision history. By Definition 2, $\zeta_{s-1}(\beta) = \zeta_{s-1}(\omega_{s-1})$.

[24]Proof. Let $\langle(\omega_0, 0), (\omega_1, 1), ..., (\omega_s, s)\rangle$ be a sequence that satisfies the properties of Definition 5. Then, since $\alpha \in B_0B_1B_2...B_{s-1}E$, $\mathcal{B}_0(\alpha) \subseteq B_1B_2...B_{s-1}E$; thus, since $\omega_1 \in \mathcal{B}_0(\alpha)$, $\omega_1 \in B_1B_2...B_{s-1}E$. Thus $\mathcal{B}_1(\omega_1) \subseteq B_2...B_{s-1}E$, etc. Thus $\mathcal{B}_{s-1}(\omega_{s-1}) \subseteq E$ and hence, since $\beta = \omega_s$ and $\omega_s \in \mathcal{B}_{s-1}(\omega_{s-1}), \beta \in E$.

[25]For example, in the game of Figure 2, $f_{BI}(\emptyset) = f_{BI}(a_1) = f_{BI}(a_1a_2) = a_1a_2a_3$, $f_{BI}(b_1) = b_1d_2$ and, for every terminal history z, $f_{BI}(z) = z$,

[26]Note that $\mathcal{B}_{m-1}(\beta) \neq \varnothing$, since, by Definition 2, $\mathcal{B}_{i,m-1}(\beta) \neq \varnothing$ and by Remark 2, $\mathcal{B}_{m-1}(\beta) = \mathcal{B}_{i,m-1}(\beta)$, where i is the active player at $\zeta_{m-1}(\beta)$.

[27]Recall also that, by Definition 2, for every action $a \in A(\zeta_{m-1}(\beta))$, there is an $\omega' \in \mathcal{B}_{m-1}(\beta)$ such that $\zeta_m(\omega') = \zeta_{m-1}(\omega')a$.

If $\zeta_{k-1}(\beta)$ is a terminal history, there is nothing to prove, since $\mathcal{B}_{k-1}(\beta) = \varnothing$. Suppose, therefore, that $\zeta_{k-1}(\beta)$ is a decision history. Let i be the active player, that is, the player who moves at $\zeta_{k-1}(\beta)$. Fix an arbitrary $\omega \in \mathcal{B}_{k-1}(\beta)$ (by Definition 2, $\mathcal{B}_{k-1}(\beta) \neq \varnothing$). Then ω is reachable from α with k steps (since, by hypothesis, β is reachable from α with $k-1$ steps). By the induction hypothesis $\zeta(\omega) = f_{BI}(\zeta_k(\omega))$. Thus (A.2) holds. Since $\alpha \in B_0 B_1 ... B_{k-2} \mathbf{T}_{k-1}$, by Remark 4 $\beta \in \mathbf{T}_{k-1}$; thus, since $\mathcal{B}_{k-1}(\beta) \neq \varnothing$,

$$\beta \in \mathcal{B}_{k-1}(\beta). \tag{A.3}$$

Thus, by (A.2),

$$\zeta(\beta) = f_{BI}(\zeta_k(\beta)). \tag{A.4}$$

Since $\alpha \in B_0 B_1 ... B_{k-2} \mathbf{R}_{k-1}$, by Remark 4 $\beta \in \mathbf{R}_{k-1}$, that is, player i is rational at state β and time $k-1$. By (A.2) at state β and time $k-1$ player i believes that after his move the play will continue according to the BI solution. Hence the action chosen by i at $\zeta_{k-1}(\beta)$ is the optimal action there given those beliefs (i.e. the action dictated by the BI solution), that is, the truncation of $f_{BI}(\zeta_{k-1}(\beta))$ at date k is equal to $\zeta_k(\beta)$:

$$\zeta_k(\beta) = f_{BI}(\zeta_{k-1}(\beta))_k. \tag{A.5}$$

It follows from (A.4) and (A.5) that $\zeta(\beta) = f_{BI}(\zeta_{k-1}(\beta))$. This completes the proof of (A.1).

Next we show that

$$\forall \omega \in \mathcal{B}_0(\alpha), \;\; \zeta(\omega) = f_{BI}(\zeta_1(\omega)). \tag{A.6}$$

Fix an arbitrary $\omega \in \mathcal{B}_0(\alpha)$. Then ω is reachable from α with 1 step and thus, by (A.1), $\zeta(\omega) = f_{BI}(\zeta_1(\omega))$. Thus the active player at state α and date 0 believes that after her move the play will continue according to the BI solution. Since $a \in \mathbf{R}_0$, it follows that the action chosen by the active player at $\zeta_0(\alpha) = \emptyset$ is the optimal action there given those beliefs, that is, the truncation of $f_{BI}(\emptyset)$ at date 1 is equal to $\zeta_1(\alpha)$:

$$\zeta_1(\alpha) = f_{BI}(\emptyset)_1. \tag{A.7}$$

Since $\mathcal{B}_0(\alpha) \neq \varnothing$ and $\alpha \in \mathbf{T}_0$, $\alpha \in \mathcal{B}_0(\alpha)$. Thus, by (A.6), $\zeta(\alpha) = f_{BI}(\zeta_1(\alpha))$. It follows from this and (A.7) then $\zeta(\alpha) = f_{BI}(\emptyset)$.

Proof of Proposition 2. Fix a perfect-information game G and define the following model of it: $\Omega = Z$ (recall that Z is the set of terminal histories), $T = \{0, 1, ..., m = \ell^{\max}\}$ (recall that $\ell^{\max}$ is the depth of the game, that is, the length of its maximal histories) and ζ is the identity function (that is, $\zeta(z) = z$, for every $z \in Z$). Let $f_{BI} : H \to Z$ be the function defined in the proof of Proposition 1. Fix an arbitrary player i, an arbitrary $z \in Z$ and an arbitrary $t \in T$. If z_t is not a decision history of player i, then we set $\mathcal{B}_{i,t}(z) = \varnothing$; if z_t is a decision history of player i then we set $\mathcal{B}_{i,t}(z) = \left\{z' \in Z : z'_t = z_t \text{ and } z' = f_{BI}(z'_{t+1})\right\}$, that is, $\mathcal{B}_{i,t}(z)$ is the set of terminal histories that (i) coincide with z up to date t and (ii) are reached by following the backward-induction choices from date $t+1$. For example, for the game of Figure 2 (whose backward-induction solution is $(a_1, a_2, (a_3, d_2))$ with corresponding terminal history $a_1 a_2 a_3$) the model just described is shown in Figure A.6.

By construction of the belief relations and by definition of backward-induction solution, at any state z and date t, if player i is active at z_t then he is rational there if and only if the action he takes there is the one prescribed by the backward-induction solution, that is, $z \in \mathbf{R}_t$ if and only

if $z_{t+1} = (f_{BI}(z_t))_{t+1}$.[28] Let $\hat{z}$ be the terminal history reached by the backward-induction solution, that is, $\hat{z} = f_{BI}(\emptyset)$. Then we have that $\hat{z} \in B_t(\hat{z})$ for every date $t \in T$ such that $B_t(\hat{z}) \neq \varnothing$ and thus $\hat{z} \in \mathbf{T}_t$ (that is, for every date t, the beliefs of the active player at $\hat{z}$ are locally correct). Thus $\hat{z} \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0(\mathbf{T}_1 \cap \mathbf{R}_1) \cap B_0B_1(\mathbf{T}_2 \cap \mathbf{R}_2) \cap ... \cap B_0B_1...B_{m-2}(\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1})$.

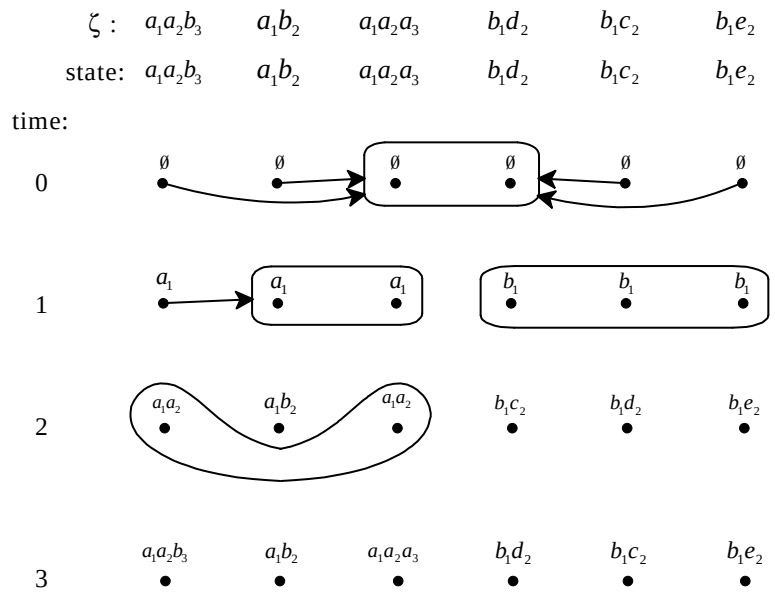


Figure A.6: The model described in the proof of Proposition 2 for the game of Figure 2.

References

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

[2] Robert Aumann. On the centipede game. *Games and Economic Behavior*, 23:97–105, 1998.

[3] Dieter Balkenborg and Eyal Winter. A necessary and sufficient epistemic condition for playing backward induction. *Journal of Mathematical Economics*, 27:325–345, 1997.

[4] Alexandru Baltag, Sonja Smets, and Jonathan Zvesper. Keep hoping for rationality: a solution to the backward induction paradox. *Synthese*, 169:301–333, 2009.

[5] Pierpaolo Battigalli and Giacomo Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53:149–225, 1999.

[6] Pierpaolo Battigalli, Alfredo Di-Tillio, and Dov Samet. Strategies and interactive beliefs in dynamic games. In Daron Acemoglu, Manuel Arellano, and Eddie Dekel, editors, *Advances in Economics and Econometrics. Theory and Applications: Tenth World Congress*. Cambridge University Press, Cambridge, 2012.

[28] In the model of Figure A.6 $\mathbf{R}_0 = \{a_1a_2b_3, a_1b_2, a_1a_2a_3\}$, $\mathbf{R}_1 = \{a_1a_2b_3, a_1a_2a_3, b_1d_2\}$ and $\mathbf{R}_2 = \{a_1b_2, a_1a_2a_3, b_1d_2, b_1c_2, b_1e_2\}$. Thus $B_0\mathbf{R}_1 = B_1\mathbf{R}_2 = B_0B_1\mathbf{R}_2 = Z$. Note that $a_1b_2 \in B_0\mathbf{R}_1$ but $a_1b_2 \notin \mathbf{R}_1$ and thus at a_1b_2 Player 1 has erroneous beliefs at date 0 about the rationality of the active player at date 1. In this model $\mathbf{T}_0 = \{a_1a_2a_3, b_1d_2\}$, $\mathbf{T}_1 = Z\backslash\{a_1a_2b_3\}$ and $\mathbf{T}_2 = Z$.

[7] Pierpaolo Battigalli and Marciano Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106:356–391, 2002.
[8] Elchanan Ben-Porath. Nash equilibrium and backwards induction in perfect information games. *Review of Economic Studies*, 64:23–46, 1997.
[9] Adam Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.
[10] Thorsten Clausing. Doxastic conditions for backward induction. *Theory and Decision*, 54:315–336, 2003.
[11] Thorsten Clausing. Belief revision in games of perfect information. *Economics and Philosophy*, 20:89–115, 2004.
[12] Yossi Feinberg. Subjective reasoning - dynamic games. *Games and Economic Behavior*, 52:54–93, 2005.
[13] Itzhak Gilboa. Can free choice be known? In Cristina Bicchieri, Richard Jeffrey, and Brian Skyrms, editors, *The logic of strategy*, pages 163–174. Oxford University Press, 1999.
[14] Carl Ginet. Can the will be caused? *The Philosophical Review*, 71:49–55, 1962.
[15] Alvin Goldman. *A theory of human action*. Princeton University Press, 1970.
[16] Joseph Halpern. Substantive rationality and backward induction. *Games and Economic Behavior*, 37:425–435, 2001.
[17] Marion Ledwig. The no probabilities for acts-principle. *Synthese*, 144:171–180, 2005.
[18] Isaac Levi. *Hard choices*. Cambridge University Press, 1986.
[19] Isaac Levi. *The covenant of reason: rationality and the commitments of thought*. Cambridge University Press, 1997.
[20] Antonio Penta. Robust dynamic mechanism design. Technical report, University of Wisconsin, Madison, 2009.
[21] Andrés Perea. Epistemic foundations for backward induction: an overview. In Johan van Benthem, Dov Gabbay, and Benedikt 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, 2007.
[22] Andrés Perea. Belief in the opponents' future rationality. Technical report, Maastricht University, August 2011.
[23] Andrés Perea. *Epistemic game theory: reasoning and choice*. Cambridge University Press, Cambridge, 2012.
[24] Antonio Quesada. From common knowledge of rationality to backward induction. *International Game Theory Review*, 5:127–137, 2003.
[25] Dov Samet. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior*, 17:230–251, 1996.
[26] George L. S. Shackle. *Time in economics*. North Holland Publishing Company, Amsterdam, 1958.
[27] Wolfgang Spohn. Where Luce and Krantz do really generalize Savage's decision model. *Erkenntnis*, 11:113–134, 1977.
[28] Wolfgang Spohn. *Strategic Rationality*, volume 24 of *Forschungsberichte der DFG-Forschergruppe Logik in der Philosophie*. Konstanz University, 1999.
[29] Robert Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.
[30] Robert Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.
[31] Robert Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128:169–199, 2006.