Bonanno, Giacomo

Working Paper

An epistemic characterization of generalized backward induction

Working Paper, No. 13-2

Provided in Cooperation with:
University of California Davis, Department of Economics

Suggested Citation: Bonanno, Giacomo (2013) : An epistemic characterization of generalized backward induction, Working Paper, No. 13-2, University of California, Department of Economics, Davis, CA

This Version is available at:
https://hdl.handle.net/10419/79676

UCDAVIS
DEPARTMENT OF ECONOMICS
Working Paper Series

An epistemic characterization of generalized backward induction

Giacomo Bona
UC Davis

March 11, 2013

Paper # 13-2

We investigate the extension of backward-induction to von Neumann extensive games (where information sets have a synchronous structure) and provide an epistemic characterization of it. Extensions of the idea of backward-induction were proposed by Penta (2009) and later by Perea (2013), who also provided an epistemic characterization in terms of the notion of common belief in future rationality. The epistemic characterization we propose, although differently formulated, is conceptually the same as Perea's and so is the generalization of backward induction. The novelty of this contribution lies in the epistemic models that we use, which are dynamic, behavioral models where strategies play no role and the only beliefs that are specified are the actual beliefs of the players at the time of choice. Thus our analysis is free of (objective or subjective) counterfactuals.

Department of Economics
One Shields Avenue
Davis, CA 95616
(530)752-0741

http://www.econ.ucdavis.edu/working_search.cfm

An epistemic characterization of generalized backward induction

Giacomo Bonanno

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

March 2013

Abstract

We investigate the extension of backward-induction to von Neumann extensive games (where information sets have a synchronous structure) and provide an epistemic characterization of it. Extensions of the idea of backward-induction were proposed by Penta (2009) and later by Perea (2013), who also provided an epistemic characterization in terms of the notion of common belief in future rationality. The epistemic characterization we propose, although differently formulated, is conceptually the same as Perea's and so is the generalization of backward induction. The novelty of this contribution lies in the epistemic models that we use, which are dynamic, behavioral models where strategies play no role and the only beliefs that are specified are the actual beliefs of the players at the time of choice. Thus our analysis is free of (objective or subjective) counterfactuals.

1 Introduction

The notion of backward induction in dynamic games with perfect information is well known and its epistemic foundations have been studied extensively.[1]

[1] For recent surveys of the literature see Brandenburger (2007), Perea (2007).

We consider the extension of the backward-induction procedure to von Neumann extensive games (where information sets have a synchronous structure; we call this procedure *generalized backward induction*) and provide an epistemic characterization of it. Neither of these two steps is new: extensions of the idea of backward-induction were proposed by Penta (2009) and later by Perea (2013), who also provided an epistemic characterization in terms of the notion of common belief in future rationality.[2] The epistemic characterization we propose, although differently formulated, is conceptually the same as Perea's and so is the generalization of backward induction. The novelty of this contribution lies in the notion of epistemic model that we use, which was first introduced in Bonanno (2013).

The epistemic models of dynamic games used in the literature[3] are static structures that postulate, for each player, a complex set of conditional belief hierarchies. In these models, at every information set of his a player holds a belief about (a) the opponents' chosen strategies, (b) the beliefs that the opponents have, at their information sets, about the other players' chosen strategies, (c) the beliefs that the opponents have, at their information sets, about the beliefs their opponents have, at their information sets, about the other players' chosen strategies, and so on. These complex structures are needed to capture intricate subjunctive conditionals such as "if I were to move across then he would believe that I am such-and-such a player, and he will believe that if he were to move across then I would move across again and consequently he would move across." (Skyrms et al. (1999), p. 276). Moreover, in the standard models, subjunctive conditionals or counterfactuals are also implicit in the use of strategies. For dynamic games with perfect information Bonanno (2013) introduced a simpler kind of models that are explicitly dynamic and make no use of (objective or subjective) counterfactuals or dispositional belief revision; furthermore, these are "behavioral" models in which strategies play no role: states are described in terms of the actual choices made by the players rather than in terms of hypothetical plans.[4] In these models there are no hypothetical beliefs or belief revision: only the actual beliefs of a player when it is her turn to move.

[2] Defined as follows: players are rational and always believe in their opponents' future rationality and believe that every opponent always believes in his opponents' future rationality and that every opponent always believes that every other player always believes in his opponents' future rationality, and so on. The first version of the paper was written in 2010 but we shall quote from the most recent version (February 2013).

[3] See, for example, Baltag et al. (2009), Battigalli and Siniscalchi (2002), Perea (2013; 2012).

[4] Behavioral models were introduced by Samet (1996).

We extend the epistemic models of Bonanno (2013) to games with imperfect information. We use a dynamic framework where the rationality of a player's choice is judged on the basis of the *actual beliefs* that she has at the time she makes that choice. The set of "possible worlds" is given by state-instant pairs (ω, t), where each state ω specifies the entire play of the game. 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 h is the decision history reached at state ω and time t and i is the active player there, then player i has to choose an action from the set $A(h)$ of available actions at h. In order to make this choice player i will form some beliefs about (1) what happened up to this point in the game (that is, which history in her information set has been reached) and (2) what will happen if she chooses action a, for every $a \in A(h)$. These beliefs are then used to assess the rationality of the choice that she ends up making at state ω. We use a very weak notion of rationality, known as "material rationality" (Aumann (1998)): at every state-instant pair (ω, t) a player is rational if (1) either she is not active there or (2) the action she ends up taking at (ω, t) is optimal given her beliefs, in the sense that it is not the case that she believes that there is another action that guarantees her a higher payoff.

The epistemic condition that we consider - which we call *forward belief of rationality* - is expressed as an event and is defined as the set of states where, at every date t, the active player (1) is rational, (2) believes that future players are rational, (3) believes that future players believe that future players are rational, (3) believes that future players believe that future players believe that future players are rational, and so on. Call this event **FBR**. We show (Proposition 1) that in an arbitrary model of a game if ω is a state such that $\omega \in$ **FBR** then the terminal history associated with ω belongs to the set of terminal histories that are the output of the generalized backward induction algorithm, which is defined as follows. Let ℓ^{max} denote the depth of the game, that is, the length of its maximal histories. The algorithm starts at information sets at depth $\ell^{max} - 1$ (these information sets are followed only by terminal histories), deletes choices that are strictly dominated there and then iterates backwards towards the root.[5] We also show (Proposition 2) that, for any game, there exists a model of it such that, for every terminal history z in the output of the algorithm there is a state ω such that $\omega \in$ **FBR** and the terminal history associated with ω is z. Thus

[5]We restrict attention to von Neumann extensive games, where information sets have a synchronous structure (decision histories that belong to the same information set have the same length). In games with perfect information and no relevant ties this algorithm yields the unique backward induction terminal history.

the notion of forward belief of rationality characterizes the (non-empty) set of terminal histories that are the output of the generalized backward induction algorithm.

Section 2 introduces the notion of dynamic, behavioral model of an extensive game, Section 3 contains the definitions of rationality and of generalized backward induction and Section 4 provides the epistemic characterization. Note that, since the word 'epistemic' refers to knowledge, while we deal with the more general notion of - possibly erroneous - belief, a better expression would be 'doxastic characterization'. Indeed, unlike the condition provided in Bonanno (2013) which involves the hypothesis of locally correct beliefs, **FBR** is completely "Truth-free" (that is, purely doxastic) and thus, as a corollary, provides an alternative characterization of backward induction in perfect information games with no relevant ties. Section 4 concludes with a discussion of of the proposed approach and of relevant literature. The proofs are given in the Appendix.

2 Models of extensive games

We shall use the history-based definition of extensive-form game (see, for example, Osborne and Rubinstein (1994)). 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 $1 \leq i \leq k$, the sequence $h' = \langle a_1, ..., a_i \rangle$ 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 without chance moves* is a tuple $\langle A, H, N, \iota, \{\approx_i\}_{i \in N} \rangle$ 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 $h \in H$, we denote by $A(h)$ the set of actions available at h, that is, $A(h) = \{a \in A : ha \in H\}$. Thus $A(h) \neq \varnothing$ if and only if $h \in D$.
- A finite set $N = \{1, ..., n\}$ 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.

- For every player $i \in N$, $\approx_i$ is an equivalence relation on D_i. The interpretation of $h \approx_i h'$ is that, when choosing an action at history $h \in D_i$, player i does not know whether she is moving at h or at h'. The equivalence class of $h \in D_i$ is denoted by $I_i(h)$ and is called an *information set of player i*; thus $I_i(h) = \{h' \in D_i : h \approx_i h'\}$. The following restriction applies: if $h' \in I_i(h)$ then $A(h') = A(h)$, that is, the set of actions available to a player is the same at any two histories that belong to the same information set of that player.[6]

Notation. If h and h' are decision histories, we write $h' \in I(h)$ as a short-hand for $h' \in I_{\iota(h)}(h)$. Thus $h' \in I(h)$ means that h and h' belong to the same information set (of the player who moves at h and h').

Given an extensive form, one obtains an *extensive game with ordinal payoffs* by adding, for every player $i \in N$, a preference relation $\succsim_i$ over the set Z of terminal histories (the interpretation of $z \succsim_i z'$ is that player i considers terminal history z to be at least as good as terminal history z'). It is customary to replace the preference ranking $\succsim_i$ with a *utility (*or *payoff) function* $U_i : Z \to \mathbb{R}$ (where $\mathbb{R}$ denotes the set of real numbers) satisfying the property that $U_i(z) \geq U_i(z')$ if and only if $z \succsim_i z'$.

Remark 1. *We will only consider ordinal payoffs and qualitative beliefs in order to highlight the novel features of our approach in as simple a framework as possible. The analysis can be extended to the case where the players' preferences are represented by von Neumann-Morgenstern utility functions and beliefs are probabilistic.*[7]

[6] It is common to impose a further requirement, known as *perfect recall*, according to which a player always remembers her own past moves. Since perfect recall is not needed for our results we are not assuming it.

[7] The traditional approach postulates that every player has a preference relation over the set of lotteries over terminal histories that satisfies the axioms of expected utility. This is not an innocuous assumption, since the game under consideration is implicitly taken to be common knowledge among the players. Thus not only is it commonly known who the players are, what choices they have available and what the possible outcomes are, but also how each player ranks those outcomes. While it is certainly reasonable to postulate that a player knows his own preferences, it is much more demanding to assume that a player knows the preferences of his opponents. If those preferences are expressed as ordinal rankings, this assumption is less troublesome than in the case where preferences also incorporate attitudes to risk (that is, the utility functions that represent those preferences are von Neumann-Morgenstern utility functions).

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 brackets and without commas: thus instead of writing $\langle \emptyset, a_1, a_2, a_3, a_4 \rangle$ we will simply write $a_1a_2a_3a_4$.

We shall restrict attention to the class of von Neumann extensive forms, which is defined as follows.[8]

Definition 2.1. An extensive form is a *von Neumann extensive form* if, for every player $i \in N$ and for every two decision histories $h, h' \in D_i$, if $h' \in I_i(h)$ (that is, h and h' belong to the same information set of player i) then $\ell(h) = \ell(h')$. Thus any two decision histories that belong to the same information set have the same length.

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*.

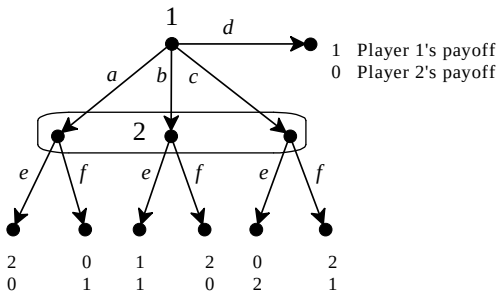
Definition 2.2. Given a von Neumann extensive form G, a *state-time representation of* G is a triple $\langle \Omega, T, \zeta \rangle$ where Ω is a set of states, $T = \{0, 1, ..., m\}$ with $m \geq \ell^{\max} - 1$ (recall that $\ell^{\max}$ is the depth of the game) and $\zeta : \Omega \to Z$ is a 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. Since the extensive form is von Neumann, whenever a player has to move she "knows the time", that is, she knows how many moves have been made so far. 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" at state ω, that

[8] Other authors impose the seemingly weaker assumption that there is an unambiguous ordering of the information sets (see, for example, Perea (2013)). However such games can be trivially transformed into von Neumann games by adding a fictitious player who always has singleton information sets and only one choice at each history assigned to him.

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 $\zeta(\omega)$].



ζ :	d	ce	ae	be	af	bf	cf
state:	α	β	γ	δ	ε	η	θ
time:							
0	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
1	d	c	a	b	a	b	c
2	d	ce	ae	be	af	bf	cf

Figure 1: An extensive-form game and a state-time representation of it.

Figure 1 shows an extensive form and a state-time representation of it. For every $\omega \in \Omega = \{\alpha, \beta, \gamma, \delta, \epsilon, \eta, \theta\}$ and $t \in T = \{0, 1, 2\}$ we have indicated the (partial) history $\zeta_t(\omega)$ (recall that $\emptyset$ denotes the empty history). For example, $\zeta_2(\alpha) = d$, $\zeta_1(\beta) = c$, etc.

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 hers if the choice she makes there is optimal given the beliefs that she holds *at the time at which she makes that choice.* These beliefs might be different from her initial beliefs about what would happen in the game and thus might be revised beliefs in light of the information she has at the moment. However, her prior beliefs are not relevant in assessing the rationality of her choice: what counts is what she believes at the time she makes the decision. 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 she will form some beliefs about (1) what happened up to this point in the game (that is, which history in her information set has been reached) and (2) what will happen if she 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.[9] 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, $\mathcal{B}_i(\omega) = \{\omega' \in \Omega : \omega \mathcal{B}_i \omega'\}$.[10] $\mathcal{B}_i$ is *serial* if $\mathcal{B}_i(\omega) \neq \emptyset$, 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.[11] 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 she 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 she believes that she believes E : $B_i E \subseteq B_i B_i E$) and euclideanness corresponds to negative introspection (if

[9] We restrict attention to qualitative beliefs since we are focusing on games with ordinal payoffs. As noted above (Remark 1), this is motivated by the desire to highlight the novelty of our approach without the more complex notation required by probabilistic beliefs and expected utility.

[10] In the economics and game theory literature the function $\mathcal{B}_i : \Omega \to 2^\Omega$ is called a *possibility correspondence* (or information correspondence). The two notions of accessibility relation and possibility correspondence are equivalent.

[11] 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".

the individual does not believe E then she believes that she does not believe E: $\neg B_i E \subseteq B_i \neg B_i E$) (for more details see Battigalli and Bonanno (1999)).

Definition 2.3. Given a von Neumann extensive form 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 2.2) 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 $\zeta_t(\omega') \in I_i(\zeta_t(\omega))$ [that is, $\zeta_t(\omega')$ belongs to the same information set as $\zeta_t(\omega)$].
   - 2.3. If $\omega' \in \mathcal{B}_{i,t}(\omega)$ then, for every $a \in A(\zeta_t(\omega'))$ there exists an $\tilde{\omega} \in \mathcal{B}_{i,t}(\omega)$ such that $\zeta_{t+1}(\tilde{\omega}) = \zeta_t(\omega')a$.

Condition 1 says that a player has trivial beliefs (that is, she believes everything) at all the state-instant pairs where she 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.[12]
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 where she has to choose an action from the set $A(h)$.
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.[13]
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 h' associated with state ω' at time t (that is, $h' = \zeta_t(\omega')$) belongs to

[12] 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). As noted above, the beliefs of inactive players are not relevant and thus there is no conceptual loss in letting those players believe everything.

[13] Note also that *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.

the same information set to which history h belongs (that is, $h' \in I_i(h)$); in other words, player i at time t knows that her information set $I_i(h)$ has been reached (although she might have erroneous beliefs concerning the history in $I_i(h)$ at which she is making her choice).
Condition 2.3 says that if player i considers it possible that she is at history h'(that is, $\omega' \in \mathcal{B}_{i,t}(\omega)$ and $h' = \zeta_t(\omega')$) then for *every* action a available at h', there is a state $\tilde{\omega}$ that player i considers possible at (ω, t) (that is, $\tilde{\omega} \in \mathcal{B}_{i,t}(\omega)$) where she takes action a at h'; that is, the truncation of $\zeta(\tilde{\omega})$ at time $t+1$ (namely $\zeta_{t+1}(\tilde{\omega})$) is equal to $h'a$ (recall that, by Condition 2.2, $h' \in I_i(h)$ where $h = \zeta_t(\omega)$). This means that, for every decision history that she considers possible and for every available action, player i has a belief about what will (or might) happen if she chooses that action at that decision history.

Remark 2. *Note that this modeling choice for 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 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.*[14] *In our approach a player's beliefs refer to the* deliberation *or* pre-choice stage*, where the player considers the consequences of all her actions, without pre-judging her subsequent decision.*[15] *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 α and β, and be labeled as rational at state α and irrational at state β, because the action she ends up taking at state α is optimal given those beliefs, while the action she ends up taking at state β is not optimal given those same beliefs.*

Figure 2 shows a von Neumann game and 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 rectangle or there is an arrow from ω to the rectangle containing ω'.[16] The relations shown in Figure 2 are those of the active players: the relation at date 0 is that of Player 1 ($\mathcal{B}_{1,0}$), the relation

[14] The role of counterfactuals in the standard approach is discussed in details in Bonanno (forthcoming)

[15] 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. Note that, while we do not endow a player with pre-knowledge of his forthcoming choice, the player is allowed to have beliefs about what choice she will make at a later time (if any). For an extensive discussion of this point see Section 4 in **?**.

[16] In other words, for any two states ω and ω' that are enclosed in a 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

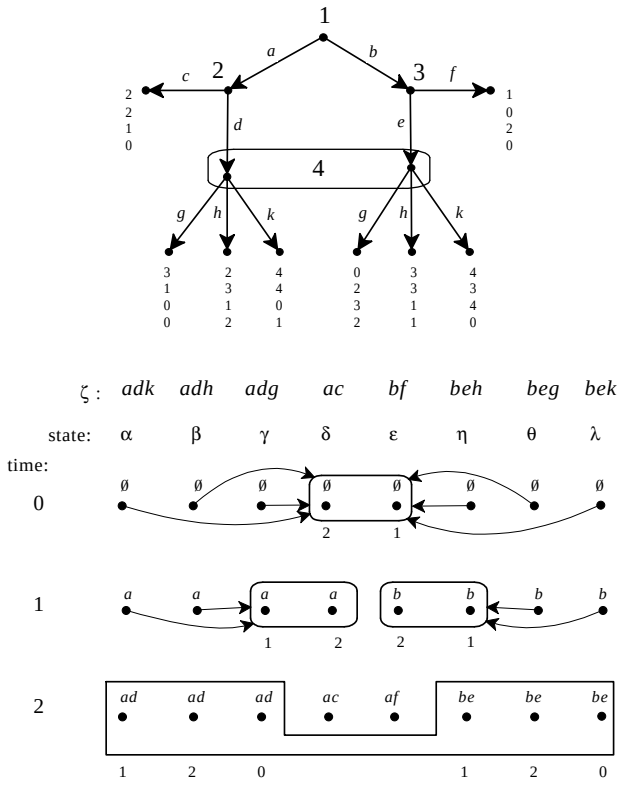


Figure 2: An von Neumann extensive game and a model of it.

at date 1 for states α, β, γ and δ is that of Player 2 ($\mathcal{B}_{2,1}$), the relation at date 1 for states ϵ, η, θ and λ is that of Player 3 ($\mathcal{B}_{3,1}$) and the relation at date 2 for states other than δ and ϵ, is that of Player 4 ($\mathcal{B}_{4,2}$).[17] Consider a state, say η. State η describes the following beliefs: at date 0 Player 1 believes - incorrectly - that if she takes action b Player 3 will follow (at date 1) with f (state ϵ) and

the rectangle) and if there is an arrow from a state ω to a rectangle then, for every ω' in the rectangle, $(\omega, \omega') \in \mathcal{B}$.

[17] Thus $\mathcal{B}_{1,0}(\omega) = \{\delta, \epsilon\}$ for every $\omega \in \Omega$, $\mathcal{B}_{2,1}(\omega) = \{\gamma, \delta\}$ for every $\omega \in \{\alpha, \beta, \gamma, \delta\}$, $\mathcal{B}_{3,1}(\omega) = \{\epsilon, \eta\}$ for every $\omega \in \{\epsilon, \eta, \theta, \lambda\}$, $\mathcal{B}_{4,2}(\omega) = \{\alpha, \beta, \gamma, \eta, \theta, \lambda\}$ for every $\omega \in \{\alpha, \beta, \gamma, \eta, \theta, \lambda\}$; for every remaining state ω, player i and date t, $\mathcal{B}_{i,t}(\omega) = \varnothing$.

she also believes that if she takes action a then Player 2 will follow (at date 1) with c (state δ); at date 1 Player 3 (knows that Player 1 played b and) believes - correctly - that if he plays e then Player 4 will follow (at date 2) with h (and if he plays f the game will end); at date 2 Player 4 considers it possible that Player 1 played a and Player 2 followed with d and also considers it possible that Player 1 played b and Player 3 followed with e. At state η Player 1 ends up playing b, Player 3 ends up playing e and Player 4 ends up playing h (while Player 2 is not active at any date). The numbers marked under the rectangles in Figure 2 are the payoffs of the active player at the relevant states.

It is worth stressing that the notion of model that we are using allows for erroneous beliefs (since the belief relations have not been assumed to be reflexive).

Remark 3. *Definition 2.3 allows for "irrational" beliefs. For example, consider a model of the game of Figure 1 where, for every $\omega \in \Omega$, $\mathcal{B}_{1,0}(\omega) = \{\alpha, \gamma, \delta, \theta\}$, capturing the following beliefs of Player 1 at time 0: "if I play a or b, Player 2 will play e, while if I play c then he will play f". Such beliefs can be considered irrational on the grounds that the choice of Player 2 cannot be influenced by what Player 1 chooses, since Player 2 does not get to observe Player 1's choice; thus a rational belief for Player 1 would require that the predicted choice(s) of Player 2 be the same, no matter what Player 1 does (provided that she gives the move to Player 2). However, this restriction on beliefs is not needed for our results and thus we do not impose it.*

3 Rationality

We shall 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 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, 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, at every history in her information set that she considers possible, her maximum utility if she takes action a is less than the minimum utility that she gets if she takes some other action b.

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 3.1. Fix a state-instant pair (ω, t) and suppose that $\zeta_t(\omega)$ is a decision history of player i. Let $a, b \in A(\zeta_t(\omega))$ be two actions available at $\zeta_t(\omega)$. We say that at (ω, t) *player i believes that b is better than a* if $\forall \omega_1, \omega_2 \in \mathcal{B}_{i,t}(\omega)$ such that $\zeta_t(\omega_1) = \zeta_t(\omega_2)$ and $\zeta_{t+1}(\omega_1) = \zeta_t(\omega_1)a$ and $\zeta_{t+1}(\omega_2) = \zeta_t(\omega_2)b$, $U_i(\zeta(\omega_1)) < U_i(\zeta(\omega_2))$ (recall that $U_i : Z \to \mathbb{R}$ is player i's utility function on the set of terminal histories).

Thus, at a decision history h of hers, player i believes that action b is better than action a if, for any history $h' \in I_i(h)$ that - according to her beliefs - might have been reached, taking action b at h' leads to terminal histories that she prefers to any terminal history that can be reached - again according to her beliefs - if she takes action a at h' [recall that, by Condition 2.3 of Definition 2.3, she must consider it possible that she takes any of her available actions at h').

Definition 3.2. Fix an arbitrary player i and an arbitrary state-instant pair (ω, t). We say that player i is *rational at* (ω, t) if and only 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 $b \in A(\zeta_t(\omega))$, it is not the case that player i believes at (ω, t) that b is better than a (see Definition 3.1).

For example, in the model of Figure 2, Player 1 is rational at date 0 and states α, β, γ and δ, because she believes that if she takes action a then her payoff will be 2 (according to her beliefs, Player 2 will follow with c) and if she takes action b her payoff will be 1 (according to her beliefs, Player 3 will follow with f) and at those states she actually ends up taking action a; Player 2 is rational at date 1 and state δ (but not at states α, β and γ); Player 3 is rational at date 1 and state ϵ (but not at states η, θ and λ) and Player 4 is rational at date 2 and every state except α and λ (because she takes action k there, which is strictly dominated by action h). Furthermore, a player is rational at any state-instant pair where she is not active (for example, Player 2 is rational at state ϵ and time 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.[18] 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

[18] By Definition 3.2 inactive players are always rational; thus $\mathbf{R}_t$ can also be described as the event that "every player is rational at date t".

history and the active player at $\zeta_t(\omega)$ is rational at (ω, t) (see Definition 3.2). Note that, in general, 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 2 we have that $\mathbf{R}_0 = \{\alpha, \beta, \gamma, \delta\}$, and $\mathbf{R}_1 = \{\delta, \epsilon\}$ and $\mathbf{R}_2 = \{\beta, \gamma, \delta, \epsilon, \eta, \theta\}$.

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 2.3, 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.

Remark 4. *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 $B_tE = \bigcap_{i \in N} B_{i,t}E$ (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 2, we have that $B_0\mathbf{R}_1 = B_0\mathbf{R}_2 = \Omega$, that is, at every state the active player at date 0 (Player 1) believes that the active player at time 1 (Player 2 at state δ and Player 3 at state ϵ) will be rational and also believes that the active player at time 2 will be rational (this is true trivially, because at states δ and ϵ there is no active player at date 2: see Definition 3.2). We also have that $B_1\mathbf{R}_2 = B_0B_1\mathbf{R}_2 = \Omega$, that is, at every state the active player at time 1 believes that the active player at time 2 will be rational and the active player at date 0 believes that the active player at date 1 believes that the active player at time 2 will be rational.

Note that the models that 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).

4 Forward belief of rationality

Fix a von Neumann extensive game and let $m = \ell^{max}$ (recall that ℓ^{max} is the depth of the game, that is, the length of the maximal histories). We shall investigate

the implications of a doxastic condition that we call *forward belief of rationality*, defined as the intersection of the following events:[19]

1. At every date the active player is rational: $\mathbf{R}_0 \cap \mathbf{R}_1 \cap \mathbf{R}_2 \cap ... \cap \mathbf{R}_{m-1}$.
2. At very date the active player believes that future players are rational: $B_0(\mathbf{R}_1 \cap \mathbf{R}_2 \cap ... \cap \mathbf{R}_{m-1}) \bigcap B_1(\mathbf{R}_2 \cap ... \cap \mathbf{R}_{m-1}) \bigcap \cdots \bigcap B_{m-2}\mathbf{R}_{m-1}$.
3. At very date the active player believes that future players believe that future players are rational: $B_0B_1(\mathbf{R}_2 \cap ... \cap \mathbf{R}_{m-1}) \bigcap B_1B_2(\mathbf{R}_3 \cap ... \cap \mathbf{R}_{m-1}) \bigcap \cdots \bigcap B_{m-3}B_{m-2}\mathbf{R}_{m-1}$.
4. At very date the active player believes that future players believe that future players believe that future players are rational: $B_0B_1B_2(\mathbf{R}_3 \cap \cdots \cap \mathbf{R}_{m-1}) \bigcap \cdots \bigcap B_{m-4}B_{m-3}B_{m-2}\mathbf{R}_{m-1}$.

... and so on, up to $B_0B_1 \ldots B_{m-2}\mathbf{R}_{m-1}$.

Remark 5. *Note that it is unnecessary to go beyond $t = m-1$, since, by Definition 3.2, for every $k \geq m$, $\mathbf{R}_k = \Omega$ and thus $B_{t_1}B_{t_2} \ldots B_{t_r}\mathbf{R}_k = \Omega$ for every sequence $\langle t_1, t_2, \ldots t_r \rangle$ in T ($r \geq 1$) with $t_r \neq k$.*

The formal definition is as follows. First, for $0 \leq k \leq m-1$ define $\mathbf{FBR}_k$ recursively by:

$$\mathbf{FBR}_{m-1} = \mathbf{R}_{m-1}, \text{ and, for } k < m-1,$$
$$\mathbf{FBR}_k = \mathbf{R}_k \cap B_k(\mathbf{FBR}_{k+1}) \cap \mathbf{FBR}_{k+1}.$$

Thus, for example, $\mathbf{FBR}_{m-2} = \mathbf{R}_{m-1} \cap B_{m-2}(\mathbf{R}_{m-1}) \cap \mathbf{R}_{m-1}$ and $\mathbf{FBR}_{m-3} = \mathbf{R}_{m-3} \cap B_{m-3}(\mathbf{R}_{m-1} \cap B_{m-2}(\mathbf{R}_{m-1}) \cap \mathbf{R}_{m-1}) \cap \mathbf{R}_{m-1} \cap B_{m-2}(\mathbf{R}_{m-1}) \cap \mathbf{R}_{m-1}$.[20]

Finally, define

$$\mathbf{FBR} = \mathbf{FBR}_0 \tag{1}$$

Example 1. In the model of Figure 2, $\mathbf{FBR} = \mathbf{R}_0 \cap \mathbf{R}_1 \cap \mathbf{R}_2 \cap B_0\mathbf{R}_1 \cap B_0\mathbf{R}_2 \cap B_1\mathbf{R}_2 \cap B_0B_1\mathbf{R}_2 = \{\delta\}$. Now consider the perfect information game and model shown in Figure 3 . Also for this game $\mathbf{FBR} = \mathbf{R}_0 \cap \mathbf{R}_1 \cap \mathbf{R}_2 \cap B_0\mathbf{R}_1 \cap B_0\mathbf{R}_2 \cap B_1\mathbf{R}_2 \cap B_0B_1\mathbf{R}_2$. In this model we have that $\mathbf{R}_0 = \Omega, \mathbf{R}_1 = \{\gamma, \delta, \epsilon\}, \mathbf{R}_2 = \{\beta, \gamma, \epsilon\}, B_0\mathbf{R}_1 = B_1\mathbf{R}_2 =$

[19] For example, when the depth of the game is 3 ($\ell^{max} = 3$), the event Forward Belief of Rationality, denoted by **FBR**, is given by $\mathbf{FBR} = \mathbf{R}_0 \cap \mathbf{R}_1 \cap \mathbf{R}_2 \cap B_0(\mathbf{R}_1 \cap \mathbf{R}_2) \cap B_1\mathbf{R}_2 \cap B_0B_1\mathbf{R}_2$ and when $\ell^{max} = 4$ $\mathbf{FBR} = \mathbf{R}_0 \cap \mathbf{R}_1 \cap \mathbf{R}_2 \cap \mathbf{R}_3 \cap B_0(\mathbf{R}_1 \cap \mathbf{R}_2 \cap \mathbf{R}_3) \cap B_1(\mathbf{R}_2 \cap \mathbf{R}_3) \cap B_2\mathbf{R}_3 \cap B_0B_1(\mathbf{R}_2 \cap \mathbf{R}_3) \cap B_0B_2\mathbf{R}_3 \cap B_1B_2\mathbf{R}_3 \cap B_0B_1B_2\mathbf{R}_3$.

[20] In the model of Figure 2, $\mathbf{FBR}_2 = \mathbf{R}_2 = \{\beta, \gamma, \delta, \epsilon, \eta, \theta\}$, $\mathbf{FBR}_1 = \{\delta, \epsilon\}$ and $\mathbf{FBR}_0 = \{\delta\}$. In the model of Figure 3, $\mathbf{FBR}_2 = \mathbf{R}_2 = \{\beta, \gamma, \epsilon\}$, $\mathbf{FBR}_1 = \{\gamma, \epsilon\}$ and $\mathbf{FBR}_0 = \varnothing$.

$B_0B_1\mathbf{R}_2 = \Omega$ but $B_0\mathbf{R}_2 = \varnothing$ and thus $\mathbf{FBR} = \varnothing$. On the other hand, if we change the model by modifying the beliefs of the root player from $\mathcal{B}_0(\omega) = \{\gamma, \delta, \epsilon\}$ to $\mathcal{B}_0(\omega) = \{\gamma, \epsilon\}$, for every $\omega \in \Omega$, (that is, we drop state δ) then $\mathbf{R}_0 = \{\alpha, \beta, \gamma, \delta\}$ and $B_0\mathbf{R}_2 = \Omega$ (while everything else remains the same), so that $\mathbf{FBR} = \{\gamma\}$. Note that $\zeta(\gamma) = a_1a_2b_3$, which is the unique backward induction terminal history. As shown in Proposition 1 below, this is not a coincidence.

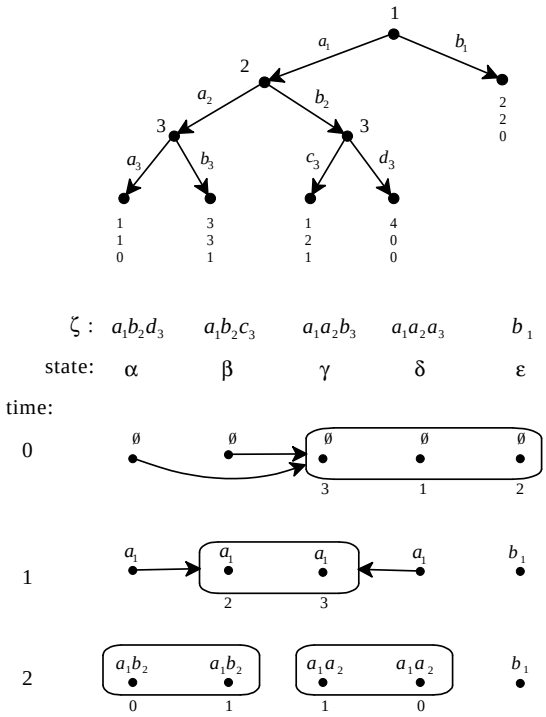


Figure 3: A perfect-information game and a model of it.

Next we introduce an algorithm that, for every von Neumann extensive-form game, selects a non-empty set of terminal histories. We call this procedure *generalized backward induction*, since it coincides with backward induction in perfect-information games with no relevant ties.[21] The procedure starts at

[21] If the output of backward induction is thought of as a terminal history rather than a strategy

information sets at depth $\ell^{max} - 1$ (these information sets are followed only by terminal histories), deletes choices that are strictly dominated there and then continues backwards towards the root. First we give an iterative definition of the set of *strictly dominated choices* at a decision history h, denoted by $D(h)$. Fix a von Neumann extensive-form game, a decision history h and let i be the player who moves at h. The set $D(h) \subseteq A(h)$ is defined recursively as follows:

1. If $\ell(h) = \ell^{max} - 1$ then $a \in D(h)$ if and only if $a \in A(h)$ and there exists a $b \in A(h)$ such that, for every $h' \in I_i(h)$, $U_i(h'a) < U_i(h'b)$ [that is, if there is another choice b which yields a higher utility than a at every history in the information set containing h; in other words, if a is strictly dominated by some other choice at $I_i(h)$].

2. Having defined $D(h)$ for every decision history h such that $\ell(h) = k$, with $0 < k \leq \ell^{max} - 1$, define $D(h)$ for a decision history h such that $\ell(h) = k - 1$ as follows: $a \in D(h)$ if and only if $a \in A(h)$ and there exists a $b \in A(h)$ such that, for every $h' \in I_i(h)$, the following holds: if $z', z'' \in Z$ are such that $z' = h'aa_1 \dots a_p$ $(p \geq 0)$ and $z'' = h'bb_1 \dots b_q$ $(q \geq 0)$ and, for all $j = 1, \dots, p$ and $k = 1, \dots, q$, $a_j \notin D(h'aa_1 \dots a_{j-1})$ and $b_k \notin D(h'bb_1 \dots b_{k-1})$ (taking $a_0 = a$ and $b_0 = b$) then $U_i(z') < U_i(z'')$ [that is, if there exists another choice that yields a higher utility than a at $I_i(h)$ assuming that only undominated actions are played after the choice at $I_i(h)$].

Next we define the following function $f_{BI} : H \to 2^Z$: (1) if $h \in Z$ then $f_{BI}(h) = \{h\}$ and (2) if h is a decision history then (defining ha_0 as h)

$$f_{BI}(h) = \{z \in Z : z = ha_1a_2...a_m \text{ and, } \forall i = 1, ..., m, \ a_i \notin D(ha_1...a_{i-1})\}.$$

Thus $f_{BI}(h)$ is the set of terminal histories that can be reached from h by following only undominated choices.

Finally define the set **BI** $\subseteq Z$ as follows:

$$\textbf{BI} = f_{BI}(\emptyset). \tag{BI}$$

Thus **BI** is the set of terminal histories that can be reached from the empty history (the root of the tree) by following only undominated choices.

Example 2. In the game of Part A of Figure 4, $D(b) = D(\emptyset) = \varnothing$ and thus $f_{BI}(b) = \{bc, bd\}$ and $BI = f_{BI}(\emptyset) = \{a, bc, bd\}$. In the game of Part B of Figure 4, $D(a) = \{c\}$,

profile.

$D(b) = \{f\}$, $D(\emptyset) = \{b\}$ and thus $f_{BI}(a) = \{ad\}$, $f_{BI}(b) = \{be\}$ and $BI = f_{BI}(\emptyset) = \{ad\}$. In the game of Part C of Figure 4, $D(bd) = D(be) = D(cd) = D(ce) = \{u\}$, $D(b) = D(c) = \{e\}$, $D(\emptyset) = \{c\}$ and thus $f_{BI}(b) = f_{BI}(bd) = \{bds, bdt\}$, $f_{BI}(be) = \{bes, bet\}$, $f_{BI}(c) = f_{BI}(cd) = \{cds, cdt\}$, $f_{BI}(ce) = \{ces, cet\}$ and $BI = f_{BI}(\emptyset) = \{a, bds, bdt\}$.[22]

Remark 6. *In a model of a game, for every state $\omega \in \Omega$ and for every date t with $0 \leq t \leq m-1$, $\zeta(\omega) \in f_{BI}(\zeta_t(\omega))$ if and only if (1) if $a \in A(\zeta_t(\omega))$ is such that $\zeta_{t+1}(\omega) = \zeta_t(\omega)a$ then $a \notin D(\zeta_t(\omega))$ and (2) $\zeta(\omega) \in f_{BI}(\zeta_{t+1}(\omega))$.*

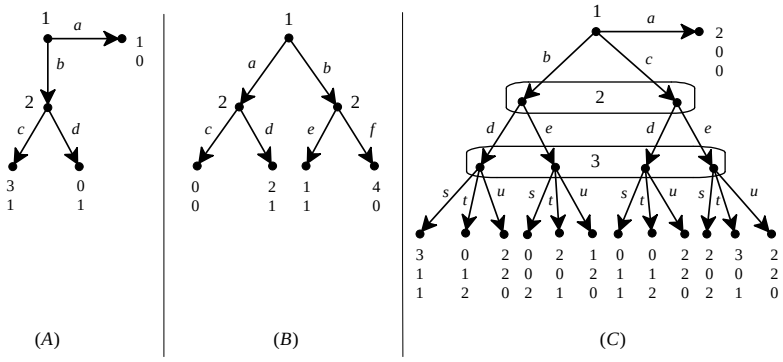


Figure 4: Three extensive games.

The following propositions state that the notion of forward belief of rationality characterizes the output of the generalized backward induction procedure. The proofs are given in the Appendix.

Proposition 1. *Fix an arbitrary von Neumann extensive game and an arbitrary model of it. Then, for every $\omega \in \Omega$, if $\omega \in$ **FBR** then $\zeta(\omega) \in$ **BI**.*

Proposition 2. *Fix an arbitrary von Neumann extensive game G. Then there exists a model of G such that, for every $z \in$ **BI**, there is a state ω such that $\omega \in$ **FBR** and $\zeta(\omega) = z$.*

[22] In the game of Figure 2, $D(ad) = D(be) = \{k\}$, $D(a) = D(b) = D(\emptyset) = \varnothing$ and thus $f_{BI}(ad) = \{adg, adh\}$, $f_{BI}(be) = \{beg, beh\}$, $f_{BI}(a) = \{ac, adg, adh\}$, $f_{BI}(b) = \{bf, beg, beh\}$, $f_{BI}(\emptyset) = \{ac, adg, adh, bf, beg, beh\}$. In the game of Figure 3, $D(a_1a_2) = \{a_3\}$, $D(a_1b_2) = \{d_3\}$, $D(a_1) = \{b_2\}$, $D(\emptyset) = \{b_1\}$ and thus $f_{BI}(\emptyset) = f_{BI}(a_1) = f_{BI}(a_1a_2) = \{a_1a_2b_3\}$ and $f_{BI}(a_1b_2) = \{a_1b_2c_3\}$.

An extensive game has a *perfect information* if and only if every information set is a singleton, that is, if $h \in D$ then $I(h) = \{h\}$. 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 $U_i(z) \neq U_i(z')$. In a perfect-information game without relevant ties **BI** is a singleton and consists of the unique terminal history that is associated with the backward-induction solution.

Corollary 1. *For perfect-information games with no relevant ties,* **FBR** *provides a doxastic characterization of the backward induction outcome.*[23]

5 Discussion

As noted in the introduction, the content of this paper is closely related to the ideas put forward in Penta (2009), Perea (2013). However the class of models used and the general philosophy is different. Below we highlight the main differences between our approach and Perea's approach in terms of the models used, the epistemic condition and the corresponding algorithm.

As noted in the introduction, the standard models used in the literature (for example, Battigalli and Siniscalchi (2002), Baltag et al. (2009), Penta (2009), Perea (2013; 2012)), are static structures where players are modeled as choosing strategies (that is, complete hypothetical plans) and are endowed with complex hierarchies of conditional beliefs. Such models incorporate a complex web of subjunctive conditionals referring to (1) the players' behavior (through strategies: "if I were to find myself at information set I then I would choose action a"), (2) the players' belief revision policies ("I do not expect that my information set I will be reached, but if it were to happen then I would have such and such beliefs") and (3) hierarchical constructions involving them ("I believe that if I were to play a then Player 2 would be surprised and would form the belief that I am of such-and-such a type and would then play b believing that I would subsequently believe that he played c and therefore I would react by playing d").

Is this complexity really necessary? The purpose of this paper was to show that the answer is negative. The models proposed here are much less demanding. First of all, strategies do not play any role in these models: a state

[23] Bonanno (2013) provides an alternative epistemic characterization of backward induction for perfect-information games in the class of models considered here, which is in terms of the beliefs of the root player and involves the hypothesis of locally correct beliefs. Thus **FBR** provides an alternative, '"Truth-free", characterization of backward induction (it can be shown that the condition given in Bonanno (2013) implies **FBR**).

only specifies *what moves are actually made* in the game and thus is silent about what players who were not called upon to move would have done if the play of the game had been different (those players might or might not have formulated hypothetical plans, but those plans are *not* part of the description of the state). Secondly, the only beliefs that are used in these models are the *beliefs of the active players at the time of choice*. No belief revision is postulated nor necessary. Thirdly, conditionals do enter into the analysis, but they are *conditionals of deliberation* for which the indicative mood seems more appropriate than the subjunctive mood (see DeRose (2010)). These conditionals are meant to capture the "exploratory" beliefs of the active player ("what will happen if I play a? what will happen if I play b?") and are modeled by taking beliefs to be *pre-choice beliefs* and thus not endowing the active player with a belief concerning what he is about to do (see Remark 2 above and the discussion of the philosophical literature on this point contained in Section 4 of Bonanno (2013)).

The fact that strategies play no role in our models reflects a different philosophy about the nature of theoretical predictions in game theory. Proposition 1 shows that the implications of a particular epistemic hypothesis is an *outcome or terminal history* not a set of strategies. To illustrate this point, consider a game where the player who moves at the root, call her Player 1, has two choices: choice a ends the game with a payoff of 2 for her, while choice b is followed by several choices of her opponents, perhaps with a very complex pattern of imperfect information; however, at every terminal history that follows choice b Player 1 gets a payoff strictly less than 2. In a model of such a game in the sense of Definition 2.3, at any state where Player 1 is rational she will end the game by playing a: *there is no attempt to obtain secondary predictions about what the other players would do, should Player 1 end up playing b*. On the other hand, Perea's notion of common belief in future rationality is much more ambitious in that it determines also a set of strategies for every other player. Indeed, while our Corollary 1 states that forward belief of rationality in an arbitrary perfect-information game with no relevant ties implies the backward-induction *outcome*, the corresponding result in Perea (2013) (namely Theorem 6.1, p. 25) states that "every player has exactly one strategy he can rationally choose under common belief in future rationality, namely his backward induction *strategy*". Thus the prediction is in terms not only of what 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.[24] It is not

[24] Aumann (1995) also derives the entire backward-induction strategy profile from the hypothesis of common knowledge of rationality [as noted in (Samet forthcoming, Footnote 4, p. 4), Aumann

clear that any game-theoretic solution concept should be so ambitious in its reach. Furthermore, there does not seem to be an obvious criterion for judging one type of counterfactual prediction as better or more reasonable than another. For example, Perea (2013) provides an example where his notion of common belief in future rationality and the notion of extensive-form rationalizability (Pearce (1984), Battigalli (1997), Battigalli and Siniscalchi (2002)) yield the same prediction in terms of outcome but different counterfactual predictions at unreached information sets.[25] Do we really need to address those counterfactuals? Do the players need to engage in such counterfactual reasoning?[26]

The generalized backward induction (GBI) algorithm proposed here is conceptually very similar to the backward dominance (BD) procedure proposed by Perea (2013).[27] The latter can be described as follows: "start with the decision problems at the end of the game, apply the procedure there until we can eliminate nothing more, then turn to decision problems that come just before, apply the procedure there until we can eliminate nothing more, and so on" ((Perea 2013, p. 24)).[28] The main difference is that, while the BD procedure operates on *strategies* and its output is *a set of strategies for each player*, the GBI procedure operates on *choices* and its output (the set $\mathbf{BI} = f_{BI}(\emptyset)$) is *a set of terminal histories*.[29] The BD procedure yields only a *superset* of the strategies that can rationally be chosen under common belief in future rationality: in

proves that common knowledge of substantive rationality implies the backward-induction strategies but *states* the weaker claim that it implies the backward-induction outcome].

[25] In Chapter 9 of Perea (2012) the author shows that every *outcome* which can be realized under extensive form rationalizability can also be realized under common belief in future rationality.

[26] In the game described above, presumably the other players will expect Player 1 to play a and yet in the standard approach they will be modeled as engaging in counterfactual speculations and hypothetical plans concerning the eventuality that Player 1 decides to play b.

[27] Related and similar procedures are the "backwards procedure"of Penta (2009) and the *iterated conditional dominance procedure* of Shimoji and Watson (1998) (see also Chen and Micali (2013)), which selects the strategies that correspond to the notion of extensive-form rationalizability (Pearce (1984), Battigalli (1997), Battigalli and Siniscalchi (2002)). For a detailed discussion of how they relate to each other see (Perea 2013, Section 7).

[28] On the other hand, the *backward rationalizability* procedure of Penta (2009) is applied not to strategies but to the conjunction of strategies and conditional belief vectors. The author uses this procedure also for games with incomplete information and applies it to issues of mechanism design and implementation.

[29] A further difference is that the BD procedure allows for the elimination of strategies that are strictly dominated by mixed strategies, while the GBI procedure does not allow the elimination of choices that are strictly dominated by mixed choices. This difference, however, is due to the fact that we only postulated ordinal payoffs and qualitative beliefs, but it would disappear if we re-formulated the problem in terms of probabilistic beliefs, von Neumann-Morgenstern utility functions and rationality as expected utility maximization.

order to get precisely those strategies it is necessary to impose common belief in Bayesian updating. In our approach Bayesian updating is not relevant and the GBI procedure yields precisely the set of *outcomes* that are compatible with the notion of forward belief of rationality.

We conclude by reiterating that the conceptual content of the notion of forward belief of rationality is the same as that of the notion of common belief in future rationality proposed by Perea (2013).[30] The main difference is in the framework used: Perea uses the standard static "type" models with conditional belief hierarchies and strategies, while we use simpler dynamic "state-space" models that do not require the use of (objective or subjective) counterfactuals.

A Proofs

Proof of Proposition 1. Fix a von Neumann extensive game and a model of it. Let $m = \ell^{max}$ be the depth of the game. First we prove that

$$\begin{array}{l}\text{For every } t \text{ with } 0 \leq t \leq m-1 \text{ and for every } \omega \in \Omega, \\ \text{if } \omega \in \mathbf{FBR}_t \text{ then } \zeta(\omega) \in f_{BI}(\zeta_t(\omega)).\end{array} \tag{2}$$

We prove this by induction.

Base step: $t = m-1$. Fix an arbitrary $\omega \in \mathbf{FBR}_{m-1} = \mathbf{R}_{m-1}$. If $\zeta_{m-1}(\omega)$ is a terminal history, then $\zeta_{m-1}(\omega) = \zeta(\omega)$ (see Definition 2.2) and, by definition of $f_{BI}(\cdot)$, $f_{BI}(\zeta(\omega)) = \{\zeta(\omega)\}$. Thus $\zeta(\omega) \in f_{BI}(\zeta_{m-1}(\omega))$. Suppose, therefore, that $\zeta_{m-1}(\omega)$ is a decision history. Let i be the active player, that is, the player who moves at $\zeta_{m-1}(\omega)$. Fix an arbitrary $\omega' \in \mathcal{B}_{m-1}(\omega)$.[31] Then, by Definition 2.3, $\zeta_{m-1}(\omega') \in I_i(\zeta_{m-1}(\omega))$. Since the depth of the game is m, after player i's move at $\zeta_{m-1}(\omega')$ the game ends and thus $\zeta_m(\omega') = \zeta(\omega')$. Since $\omega \in \mathbf{R}_{m-1}$, that is, player i is rational at state ω and time $m-1$, the choice made by player i at state ω and time $m-1$ is not strictly dominated at the information set containing $\zeta_{m-1}(\omega)$, that is, if $\zeta(\omega) = \zeta_{m-1}(\omega)a$ then $a \notin D(\zeta_{m-1}(\omega))$ and thus, by definition of $f_{BI}(\cdot)$, $\zeta(\omega) \in f_{BI}(\zeta_{m-1}(\omega))$.

Induction step: suppose that (2) 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 state β and suppose that

$$\beta \in \mathbf{FBR}_{k-1} = \mathbf{R}_{k-1} \cap B_{k-1}\mathbf{FBR}_k \cap \mathbf{FBR}_k. \tag{3}$$

[30] A related notion is that of sequential rationalizability (Asheim and Perea (2005), Dekel et al. (1999; 2002)). For a detailed discussion of how they relate to each other see (Perea 2013, Section 7).

[31] Note that $\mathcal{B}_{m-1}(\omega) \neq \emptyset$, since, by Definition 2.3, $\mathcal{B}_{i,m-1}(\omega) \neq \emptyset$ and by Remark 4, $\mathcal{B}_{m-1}(\omega) = \mathcal{B}_{i,m-1}(\omega)$.

If $\zeta_{k-1}(\beta)$ is a terminal history, then $\zeta_{k-1}(\beta) = \zeta(\beta)$ and, by definition of $f_{BI}(\cdot)$, $f_{BI}(\zeta(\beta)) = \{\zeta(\beta)\}$, so that $\zeta(\beta) \in f_{BI}(\zeta_{k-1}(\beta))$. 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 2.3, $\mathcal{B}_{k-1}(\beta) \neq \varnothing$). Since, by (3), $\beta \in B_{k-1}\mathbf{FBR}_k$ (that is, $\mathcal{B}_{k-1}(\beta) \subseteq \mathbf{FBR}_k$), $\omega \in \mathbf{FBR}_k$. Hence, by the induction hypothesis, $\zeta(\omega) \in f_{BI}(\zeta_k(\omega))$. Thus at state β and time $k-1$ player i believes that after every choice at her information set $I_i(\zeta_{k-1}(\beta))$ only terminal histories selected by the function f_{BI} can be reached. Since, by (3), $\beta \in \mathbf{R}_{k-1}$, it follows that the choice made by player i at state β and time $k-1$ is not strictly dominated conditional on the belief that future choices by the future players (if any) are not strictly dominated, that is, if $\zeta_k(\beta) = \zeta_{k-1}(\beta)a$ then $a \notin D(\zeta_{k-1}(\beta))$. By (3) $\beta \in \mathbf{FBR}_k$ and thus, by the induction hypothesis, $\zeta(\beta) \in f_{BI}(\zeta_k(\beta))$.[32] Hence $\zeta(\beta) \in f_{BI}(\zeta_{k-1}(\beta))$ (see Remark 6). This completes the proof of (2).
Now fix an arbitrary state α and suppose that $\alpha \in \mathbf{FBR}$. We need to show that $\zeta(\alpha) \in \mathbf{BI} = f_{BI}(\emptyset)$. But this is an immediate consequence of (2), since $\mathbf{FBR} = \mathbf{FBR_0}$ and $\zeta_0(\alpha) = \emptyset$.

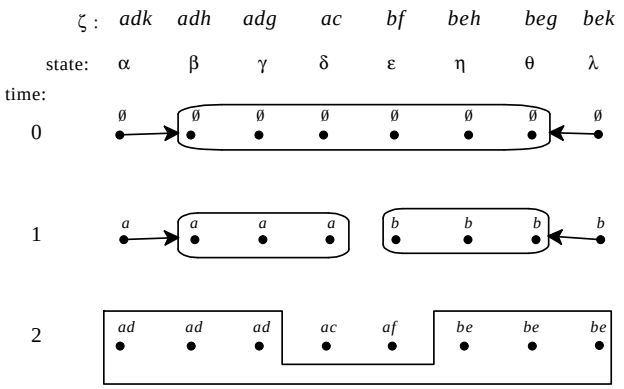


Figure 5: The model described in the proof of Proposition 2 for the game of Figure 2.

Proof of Proposition 2. Fix a von Neumann extensive-form game and define

[32]Note that this last step is crucial, since it is possible that $\beta \notin \mathcal{B}_{k-1}(\beta)$. For example, in the model shown in Figure 3, we have that $\delta \in \mathbf{R_1} \cap B_1\mathbf{FRB_2} = \mathbf{R_1} \cap B_1\mathbf{R}_2 = \{\gamma, \delta, \epsilon\} \cap \Omega = \{\gamma, \delta, \epsilon\}$ but $\delta \notin \mathbf{FBR_2} = \mathbf{R}_2 = \{\beta, \gamma, \epsilon\}$ and indeed $\zeta(\delta) = a_1a_2a_3 \notin f_{BI}(\zeta_1(\delta)) = f_{BI}(a_1) = \{a_1a_2b_3\}$.

the following model of it: $\Omega = Z$ (recall that Z is the set of terminal histories), $T = \{0, 1, ..., m = \ell^{\max} - 1\}$ (recall that $\ell^{\max}$ is the depth of the game) and ζ is the identity function (that is, $\zeta(z) = z$, for every $z \in Z$). Fix an arbitrary (z, t). If z_t is a terminal history set $\mathcal{B}_{j,t}(z) = \varnothing$ for every player $j \in N$. If z_t is a decision history of player i set $\mathcal{B}_{j,t}(z) = \varnothing$ for every player $j \neq i$ and define $\mathcal{B}_{i,t}(z)$ as follows: $z' \in \mathcal{B}_{i,t}(z)$ if and only if (1) $z'_t = I_i(z_t)$ and (2) $z' \in f_{BI}(z'_{t+1})$. Figure 5 shows the model just described for the game of Figure 2. By construction of the belief relations, 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 following holds: if a is the action at z_t such that $z_{t+1} = z_t a$ then $a \notin D(z_t)$. Now fix an arbitrary $z \in \mathbf{BI}$. Then, by construction, for every $t \in T$, $z \in f_{BI}(z_t)$, so that $z \in \mathbf{FBR}_t$. Hence $z \in \mathbf{FBR}_0 = \mathbf{FBR}$.

References

G. Asheim and A. Perea. Sequential and quasi-perfect rationalizability in extensive games. *Games and Economic Behavior*, 53:15–42, 2005.

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

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

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

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

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

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*. Cambridge University Press, Cambridge, 2013.

G. Bonanno. A dynamic epistemic characterization of backward induction without counterfactuals. *Games and Economics Behavior*, 78:31–45, 2013.

G. Bonanno. Reasoning about strategies and rational play in dynamic games. In J. van Benthem, S. Ghosh, and R. Verbrugge, editors, *Modeling strategic reasoning*, Texts in Logic and Games. Springer, forthcoming.

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

J. Chen and S. Micali. The order independence of iterated dominance in extensive games. *Theoretical Economics*, 8:125–163, 2013.

E. Dekel, D. Fudenberg, and D. Levine. Payoff information and self-confirming equilibrium. *Journal of Economic Theory*, 89:165–185, 1999.

E. Dekel, D. Fudenberg, and D. Levine. Subjective uncertainty over behavior strategies: a correction. *Journal of Economic Theory*, 104:473–478, 2002.

K. DeRose. The conditionals of deliberation. *Mind*, 119:1–42, 2010.

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

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

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, 2007.

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

A. Perea. Belief in the opponents' future rationality. Technical report, Maastricht University, February 2013. URL http://www.personeel.unimaas.nl/a.perea/Papers/FutureRat.pdf.

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*, forthcoming.

M. Shimoji and J. Watson. Conditional dominance, rationalizability, and game forms. *International of Economic Theory*, pages 161–195, 1998.

B. Skyrms, G. D. Bell, and P. Woodruff. Theories of counterfactual and subjunctive conditionals in contexts of strategic interaction. *Research in Economics*, 53:275–291, 1999.